Chapter 1
The Role of Evaluation in Improving Education

Orienting Questions

1. What is the difference between formal and informal evaluation?
2. What is the *goal* of evaluation? What *roles* can evaluation play in education?
3. What is an evaluation object? What are some examples of important evaluation objects in education?

In most advanced nations, education is increasingly viewed as a primary means for solving social problems. Indeed, in some cases, the future welfare of nations has been placed squarely on the shoulders of the schools and universities. In the United States, for example, the National Commission on Excellence in Education was created to investigate "the widespread public perception that something is seriously remiss in our educational system," and the "support of all who care about our future" was elicited to aid the commission's work (National Commission on Excellence in Education, 1983, p. 1). The first essential message in the commission's public report, entitled *A Nation At Risk: The Imperative for Educational Reform*, is that the United States' once unchallenged lead in commerce, industry, science, and technological innovation is being overtaken by competitors throughout the world. The commission's panel of distinguished leaders warned that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a nation and a people" (National Commission on Excellence in Education, 1983, p. 5). This report has sparked a national debate on education that echoes similar (if less dramatic) dialogues in countries everywhere.

Admittedly, critics of existing educational systems often overstate their case, focusing so much on the inadequacies of schools that they breed pessimism about the possibility of genuine educational improvement. But even if such pessimism is unjustified, one cannot discount the critics' concerns, for they serve to highlight one key deficit in most educational systems: the lack of effective evaluation. Without careful, systematic inquiry into the effectiveness of either current school

practices or new programs, many changes occurring in education become little more than random adoption of faddish innovations. Probably the greatest contri— butors to this inadequate evaluation are (1) the lack of dependable information about the performance of educational products, practices, and programs; and (2) the absence of established systems for producing such information.

Though it is just one step toward educational improvement, evaluation holds greater promise than any other approach in providing educators with information they need to help improve educational practices. Recognition of this fact has encouraged many educational and governmental leaders to support evaluation, and **most** educated publics agree that school programs should be evaluated. Parents want information about the curricula and teaching methods used to instruct their children. Other citizen groups want to know what results are being achieved through schools' expenditures of public funds. Because evaluation can help provide this information, lawmakers often use evaluation mandates as a means of legislating school improvement, and school and university officials accept evaluation as a necessary condition for obtaining funds for many educational programs. Many teachers and administrators scan evaluation reports to catch a clue about how well they arc doing. In short, evaluation has gained widespread acceptance in education and related fields

INFORMAL VERSUS FORMAL EVALUATION

Evaluation is not a new concept. One dictionary definition of evaluation is, " To determine the worth of: to appraise'' (*Webster's* \'t'li' LI *'end Dii'tionary,* 1960, P. 26). Given such broad focus for the term, it can be argued that evaluation has been with us always and that everyone is, in his or her own way, an evaluator. When the English adopted and improved upon the Welsh longbow, **it** was because the English yeomen saw its advantages over their own crossbows. The longbow could send an arrow through the stoutest armor and was capable of launching three arrows while the crossbow was sending one, in short, the English evaluated the longhows value for their purposes, deciding that its use would strengthen them in their struggles with the French. So they abandoned the crossbow and perfected the loiigbow, and the English armies proved invincible during most of the Hundred Years' War. By contrast, French archers experimented briefiv with the longbow, then went back to the crossbow—and continued to lose battles. Such are the perils **of poor** evaluation. Unfortunately, the faulty judgment that led the French to persist in using an inferior weapon represents an informal evaluation pattern that has been repeated throughout history.

Consider the fifth—grade teacher who decides to continue using outdated, pho.nerically inaccurate, and culturally insensitive reading hooks in her class rather than the highly regarded, up—to—date, linguistically correct, and culturally sensitive reading curriculum adopted by the school district. Her decision is most probably based **upon a** highly informal appraisal of the value of the two alternative hooks to icr iiistrttctioiial program. Or, consider the admiiiistr,iors who establish a grad— uate progralli to train professional personnel and then, without collecting any

data about the program, **Vute** to terminate it before the first graduating class has taken jobs. They too are engaged in evaluation of a sort.

Evaluation then is a basic frm of human behavior. Sometimes **it** is thorough. structured, and formal. More often it is impressionistic and private. informal evaluation occurs whenever one chooses from among available alternatives—and sometimes informality is the only practical approach. (in choosing an entrée from a dinner menu, only the most compulsive individual would conduct exit interviews with restaurant patrons to gather data in support of his dinner choice.) This informal type of evaluation, choices based on highly subjective *perceptions* of which alternative is best, is not of concern in this book. Our focus is the more formal and systematic approach to evaluation, where choices are based on *sysIenuirit* efforts to define criteria and obtain *accurate* information about alternatives (thus enabling the real value of the alternatives to he determined).

EVALUATION'S ROLES AND GOALS

Formal evaluation studies have played many roles in education, including the folio wing:

1. To provide a basis for decision making and policy formation
2. To assess student achievement
3. To evaluate curricula
4. To accredit schools
5. To monitor expenditure of public funds
6. To improve educational materials and programs.

Scriven (1973) notes that evaluation plays many roles in education, even though it has a single goal: to determine the worth or merit of whatever is being evaluated. He made the important distinction that the *goal* of evaluation is to provide answers to significant evaluative questions that are posed, whereas evaluation *roles* refer to the ways in which those answers are used. The goal usually relates to value questions, requires judgments of worth or merit, and is conceptually distinct from roles. Scriven made the distinction this way:

In terms of goals, we may say that evaluation attempts to answer certain *types of question* about certain *enmuies*. The entities are the various educational instruments (processes, personnel, procedures, programs, etc.). The types of question include questions of the form: *How well* does this instrument perform (with respect to such—and—such criteria)? Does it perform *better* than this other Instrument; U7ia merits, or drawbacks does this instrument have (i.e.. what variables from the group in which we are interested are significantly affected b its application)? Is the use ofthis instrument *north* what its costing?

the *roles* which evaluation has in a particular educational context may be enormously various: it may form part of a teacher training activity, of the process of curriculum development, of a field experiment connected with the improvement of learning theory, of an investigation preliminary to a decision about purchase or rejection of materials; **it** may be a data—gathering activity for supporting a request for tax increases or research.support. or a prelimmarv to the reward or punishment of people as in an executive training program. a prison, or a classroom. Failure ro make this rather obviom

distinction between the roles and goals of evaluation is one of the factors that has led to the dilution of what is called evaluation to the point where **it** can no longer answer the questions which are its principal goal, questions about real merit or worth. (Scriven, 1973, pp. 61—62)

We shall discuss evaluation's goal further in later chapters. Here we deal more with the roles evaluation plays.

Many authors have attempted to categorize the purposes for which evaluations are conducted. For example, Brophy, Grotelueschen, and Gooler (1974) outlined three major reasons for conducting evaluations:

1. Planning procedures, programs, and/or products
2. Improving existing procedures, programs, and/or products
3. Justifying (or not justifying) existing or planned procedures, programs, and! or products.

Most educators agree that evaluation can serve either a *formatiiie* purpose (such as helping to improve a mathematics curriculum) or a *sumtnatii'e* purpose (such as deciding whether that curriculum should be continued).' Anderson and Ball (1978) further describe the capabilities of evaluation, as applied to formal programs, in terms of six major purposes (which are not necessarily mutually exclusive):

1. To contribute to decisions about program installation
2. To contribute to decisions abouL program continuation, expansion. or certification
3. To contribute to decisions anout program modifications
4. To obtain evidence to rally support for a program
5. To obtain evidence to rally opposition to a program
6. To contribute to the understanding of basic psychological, social, and other processes. **2**

It is not our purpose to provide an exhaustive list of all the purposes educational evaluation can serve. There are many. And the list continues to grow as more educators gain experience in using evaluation fnr their benefit. Support for continued use and improvement of evaluation generally rests on one of the following arguments:

1. There is a need to plan and carry out school improvcinents in a systematic way that includes (a) identifying needs, (b) selecting the best strategies from among known alternatives, (c) monitoring changes as they occur, and (d) measuring the impact of these changes. Only through this process can educators mininiize the chance of misdirected or inconsequential changes and justify expenditures associated with beneficial changes (Stuffiebeam **&** Shinkficld, 1985). Hammond (I 973) also made a similar logical argument by stressing the use of systematic evaluation to avoid faddism, overreaction to political pressure, pendulum swinging, reliance on persuasive claims of advocates and salesmen, and resistance to information sharing (that is, a reluctance or simple lack of effort to let staff and the public know what is happening in the school). Alkin, Daillak, and White (1979) documented the use of evaluation to direct decision making about new programs.
2. There is a need for cost—benefit analysis of programs and practices that

require large amounts of money (Madaus, Airasian, & Kcllaghan, 1980). A push for accountability in public education (Lessinger & Tyler, 1971: l3rowdcr, Atkins, & Kaya, 1973; Webster. 1977) is consistent with this line of reasoning.

3. There is a need to test a number of popular theories (myths?) about the effects of education on student development. Professional experience currently dictates most teaching and school management practices. Yet **it** seems quite appropriate to ask that experience—based educational decisions be justified and supported. As Cronbach and others (1980, p. 38) noted, "The need for systematic and often subtle information to supplant or confirm casual observations is what generates the call for evaluation."

4. Educators have a professional responsibility to appraise the quality of their school programs, and they constantly seek ways of improving that quality (Spencer, 1964; Cronbach and others. 1980).

5. There is a need to reduce uncertainty about educational practices when experience is limited (Patton and others, 1978: Kennedy, Apling, & Neumann, 1980).

6. There is a need to satisfy external agencies' demands for reports, to legitimize decisions, or to improve public relations through credible, data-based decision makmg (King, Thompson, & Pechman, 1982).

Evaluation as Political Activity

An example of widely diverse roles evaluation may play is **found in the urging** of Cronbach and his colleagues (1980, p. 152) that an evaluator go beyond his or her traditional technical, scientific roles to play "...an active role in political events, preferably as a multipartisan who serves the general interest." Thoughtful evaluators have come to realize that evaluation is not just a technical procedure involving instrument development, data collection. and data analysis—it is also a political activity. Information is power, and those who possess information unavailable to others have an advantage. Moreover, whenever the special interests of different individuals or groups are being considered in an evaluation, as they are most of the time, there is opportunity for one point of view to dominate. Attempts to influence the outcomes of an evaluation, or to avoid evaluation altogether, are yet another way in which political influence interplays with evaluation. Brickell (1978), Cronbach and others (1980), House (1973, 1980), and the Joint Committee on Standards for Educational Evaluation (1981) have all examined ways in which evaluation and politics interrelate. This topic will be discussed in further detail in Chapter 17.

THE OBJECTS OF FORMAL EVALUATION STUDIES

Formal evaluation studies have been conducted to answer questions about a wide variety of educational entities, which we shall refer to here as ***eualuation** objects*. The evaluation object is whatever is being evaluated, Important evaluation objects in education include the following:

- Student development and performance
- Educator qualifications and performance
- Curriculum design and processes
- School organizational structure
- Textbooks and other curriculum materials and products
- Funded or unfunded projects
- Any aspect of school operations (school transportation, food services, health services)
- School budgets, business and finance
- Facilities, media and libraries, equipment
- Educational policies
- School-community relations
- Parent involvement in schools
- School climate
- Ideas, plans, and objectives.

Each **of** these objects may be analyzed by looking at its component parts, their interactions, and their performance within particular contexts. Objects change over time and within different contexts. The object that is described in one setting or in one time frame may be very different from that same object described elsewhere or at a different time. Descriptions of a program presented in a plan or public relations document, for example. niav or may not correspond to reality. Evaluators have found that firsthand experience with the object of the evaluation is the best way to gain valid knowledge about it.3

Whatever the evaluation object, the process requires retlcction and sensitivity to the values of others if the evaluation is to be adequate Questions like " What makes a **good** teacher?" or " What makes a good high school curriculum?" do not have easy answers. The answers are dependent on the students who are being served by the school, on available resources, on conceptions of what schools are tr (see Goodlad, 1979, for a provocative discussion ot this question), on the research literature that reports relationships between teachim variables or curriculum variables and student development, on accreditation requirements. and certainly on the values of the constituents of the school. But while simple answers do not **exist,** the benefits to students and eventually to society of addressing these important questions cannot be underestimated.

Potential and Limitations of Educational Evaluation

The usefulness of educational evaluation has led sonic persons to look to **it** as a panacea **for all** the i11s of education. But evaluation alone cannot solve all of' education's problems. **One** of the biggest mistakes 0f evaluators in the 1970s was to promise results that could not possibly be attained. Stake (1982, p. 58) noted this when he said, "Evaluator promises often leap beyond what the proposer has previously accomplished and also beyond the attainment of anyone in the held. Even ardent supporters of evaluation are forced to admit that many evaluation

studies fail to lead to significant improvements in school programs. Why? Partly it's a question of grave adequacies *in* the conceptualization and conduct of many educational evaluations, it's also a question of understanding too little about other factors that affect the usc of cvaluarion information, even from studies that **arc** well conceptualized and well conducted. In addition, we may have been limited by our unfortunate tendency to view evaluation as a series of discrete studies, rather than a continuing system of self—renewal.

A few poorly planned, badly executed, or inappropriately ignored evaluations should not surprise us; such thilings occur in every field of human endeavor. The real problem is one of frequency and significance. So many key evaluations have been disappointing or have made such little impact that even some evaluation advocates have expressed reservations about evaluation living up to **its** high potential. Indeed, unless evaluation practices improve significantly in the years ahead, its potential for improving education may never be realized. That **need** not happen. This book is intended to help educational evaluators, and those ***who*** use their results to improve the practice and utility of evaluation in the field of education.

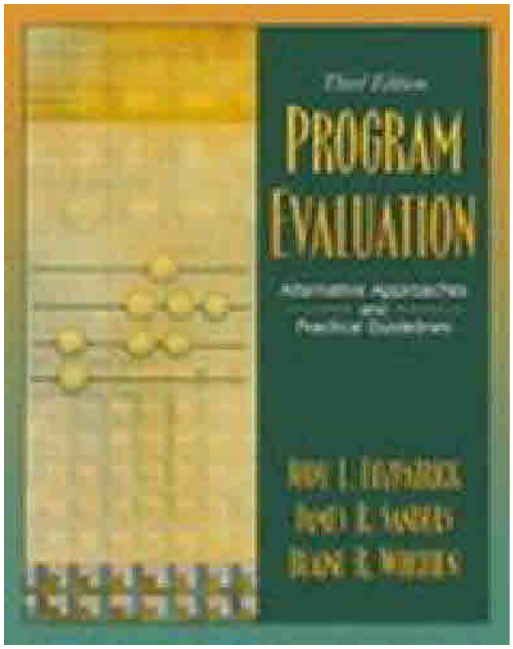
A parallel problem exists when those served by evaluation naively assume that its magic wand need only be waved over an educational enterprise to correct all its malfunctions and inadequacies. As 1—louse, Mathison, Pearsol, and Preskill (1982, p. 27) put it, "The expectations of what evaluation can accomplish ***are*** **often** beyond common—sense limits. "4 Though evaluation can be enormously useful, it is generally counterproductive for evaluators or those who depend on their work to propose evaluation as the ultimate solution to every problem or, indeed, as any sort of solution, because evaluation in and of itself wod't effect a solution—though it might suggest one. Evaluation serves to identify strengths and weaknesses, highlight the good, and expose the faulty, but not to ***correct*** problems, **for** that is the separate step of using evaluation findings.

Evaluation has a role to play in enlightening its consumers, and it **may be used** for many purposes in education. But it is only one of many influences **on** educational policies, practices, and decisions. Both its limitations and its benefits must be acknowledged. The remainder of this book is devoted to enhancing **an** understanding of how evaluation may reach its full potential as a force for improving education.

APPLICATION EXERCISES

1. Discuss the potential and limitations of educational evaluation. **Identify some** things evaluation cannot do for educators.

2. Within your own institution (if you are a university student, you might choose your university), identify several evaluation objects that you believe would be appropriate for study. For each, identify: (1) the *role* the evaluation study would play, and (2) the *,ea1* of the evaluation.


Third Edition
PROGRAM
EVALUATION
Alternative Approaches
and
Practical Guidelines

Chapter 2
The History of Evaluation in Education

Orienting Questions

1. Prior to 1920, what use was made of evaluation in the field of education? Which American educators of that period used evaluation? How did they use it?
2. Who were the major contributors to the development of educational evaluation between 1920 and 1965? What did each of these individuals contribute?
3. What major political events occurred in the late 1950s and early 1960s that greatly accelerated the growth of evaluation thought?
4. How would you characterize the growth of educational evaluation from 1965 to the present? What are some significant developments that have occurred in the field of evaluation during this period?

ormal evaluation (at least of educational and social programs) has only begun to mature as a conscious, purposeful activity, with major developments occurring over the past 20 years. As noted by some analysts:

Evaluation, as an established field, is now in its late adolescent years. The bubbling, exciting, fast_developing childhood years of the late 1960s and early 1970s gave way in the mid to late 1970s to the less self—assured, serious, introspective early adolescent years. Now, in the early 1980s, evaluation is making the transition from late adolescence to adulthood. (Conner, Altman, & Jackson, 1984, p. 13)

Yet we should not overlook the forms of evaluation that have served education prior to this recent growth spurt.

THE HISTORY AND INFLUENCE OF EVALUATION IN EDUCATION

Assuming that decisions have always been a part of education, we cn safely claim that evaluation has always had a role to play. Informal evaluation, or the way in

which people form impressions or perceptions about which educational alternatives are best, is as much a part of education as teaching itself. Yet even formal evaluation, or the use of accurate information and criteria to assign values and justify value judgments, has a much longer and more distinguished history than is generally recognized.

Educational Evaluation: Early History to A.D. 1920

The practice of evaluating individual performance was evident as early as 2000 B.C., when Chinese officials conducted civil service examinations to measure proficiency of public officials. Greek teachers, such as Socrates, used verbally mediated evaluation as part of the learning process. But formal evaluations of educational and social programs were almost nonexistent until the mid—nineteenth century.

Travers (1983) has established that prior to the mid—1800s there was little that could be construed as formal evaluation in American education.D Prior to 1837, political and religious beliefs dictated most educational choices. Communities were happy to attract and hold teachers, regardless of their competence, and if a teacher did prove incompetent in those days, formal evaluation was relatively pointless anyway—the school just closed for lack of students.

Americans Henry Barnard and Horace Mann—and later. William Torrey Harris—apparently introduced the practice of collecting data on which to base educational decisions. Their work began in the state education departments of Massachusetts and Connecticut and was continued in the United States Education Bureau with a process for collecting information to assist and support decision making. Thus were sown the seeds of educational evaluation in the United States.

During the period 1838 to 1850, Horace Mann submitted 12 annual reports to the Board of Education of the Commonwealth of Massachusetts. As Travers (1983) reported in detail, these reports identiFied all current educational concerns. with empirical support. Many of the concerns reported in those annual reports are still relevant today:

- Geographic distribution of schools
- Adequacy of outside supervision
- Financial support tor poor students who want ro attend school
- Low interest in education among community inem hers
- School finance
- Teacher competency
- Selection or construction of appropriate curriculum materials
- Adequacy of school libraries in rural areas
- Consolidation of small schools
- Teacher training
- l)iscipline
- Economic beneFits in free public education.

In 1845, the Boston School Committee undertook what became known as the

*Boston **Suri'cy,*** the first use of printed tests fiar widescale assessment of student achievement. A sample of Boston students were tested in dehiutions, geoiraphv, grammar, civil history, iiatur;iI philosophy, astronomy, writing, and arithmetic. lntcrestingly, the committee was shocked by the low level of performance in 1845 and again in 1846, and discontinued the testing in I 847 because no use was made of the results (Travers, 1983). This precursor to modern—day debate over the value of school testing was the first attempt at objectively measuring student achievement to assess the quality of a large school system.

Later, during the period of 1895— 191)5, Joseph Rice organized a similar assess— merit program carried out in a number of large school sYstems throughout the United States. l3ccause Rice had a reputation as an educational critic, he feared his criticisms might be dismissed as the grurublings of one disgruntled observer. Hence, he set out to document his claims that schooltime was inefficiently used. In his tests of spelling, for example, he found negligible differences in students' performance from one school to another, regardless of the aillount of time spent on spelling instruction. He used these data to support his proposals for restructuring spelling instruction. F us tests of arithmetic, on the other hand, revealed large differences among schools; consequently, Rice proposed the setting up of standardized examinations (Travers, 1983).

One more interesting contribution by Rice to the field of school evaluation might be noted. In 1915, he published a book entitled *The People's Gor'ernment: Efficient, Bossless, Gra/Hess*. In **it** he proposed a system for resolving controversial policy issues; namely, to bring together all relevant facts and present them to a qualified panel ofjudges. This process, proposed as a means of eliminating graft and waste in government, was an early and little known form of what later emerged as the advocate—adversary or judicial approach to evaluation.

In the early 1900s, Edward Lee Thorndike, called the father of the educational testing movement, helped persuade educators that measuring human change was worthwhile. Measurement technology for determining human abilities flourished in the United States during the first two decades of the present century. and testing emerged as the primary means of evaluating schools. By World War I, somewhere between 18 (Travers, 1983) and 40 (Madaus, Airasian, & Kellaghan, 1980) large school systems had bureaus of school research working on large—scale assessments of student achievement. It was reported that these surveys were used "for a variety of purposes: to diagiose specific system weaknesses, to standardize curricular practice, to evaluate experiments, and to assess the overall performance of a system as well as to make decisions about individuals" (Madaus, Airasian, & Kellaghan, 1980, p. 6).

The testing movement was in full swing by 1918, with individual and group tests being developed for use in many educational and psychological decisions. The Army Alpha (for literates) and Beta (for illiterates) tests, developed and used during World War I, lent credibility to the notion that good decisions about individuals could be made only when objective test information was available. Though the early school system surveys had relied mainly on criterion-referenced

tests to gather group information in school subject areas, the 1920s saw the emergence of norm—referenced tests developed for use in measuring individual performance levels.
One other event of interest occurred right before World War I. This was the commissioned evaluation of the Gary, Indiana, public school system by Abraham Flexner, beginning in 1915. The Gary plan was an innovative means of meeting the educational needs of the Gary community. This evaluation was apparently requested by the school superintendent, William Wirt, who was convinced that his schools were the best in the country and who wanted an outside group of individuals to conduct a study of the school system that would prove him right. The evaluation study was completed in 1918, at considerable cost. The conclusions stated that the Gary students were academically ipferior to comparison students, but some commentators (for example, Travers, 1983) believe the study was highly biased against the Gary plan. Results aside, this evaluation is the first evidence that we have found of a school board hiring an outside, independent team of evaluators to study and report on a controversial educational program. The political and technical problems inherent in this evaluation stud still plague educational evaluations today. Perhaps if the concepts of meta—cvaluarion and standards for evaluation had been developed back in 1915, the Gary school system might have benefited rather than suffered. After reading an account of the Gary evaluation. one cannot avoid appreciating the way in which educational evaluation has developed during the twentieth century.

Educational Evaluation: 1920—1965

The testing movement continued to Hourish during the l92s and I 93hs. with the New York Board of Regents examination appearing in 1927 and the Iowa tests in 1929. By the mid—193{)s, over one—half of the United Stares had some form of statewide testing. The development of standardized achievement tests tor USC in large—scale testing programs was a natural outgrowth of this trend. In addition. teacher—made achievement tests burgeoned, forming a basis tor most school grading systems. Personality and interest profiles were also developed during this period. The military and private industry began using these new tools to evaluate recruits or applicants for personnel selection and classification. During this period, measurement and evaluation were regarded as near] snon and the term
valuation was most often used to mean the assigning of grades or summarizing of student performance on tests. The concept of evaluation, as we know it today, was still evolving.
The 1920s also saw the emergence of the empirically minded school superintendent, embodied in Carleton Washburne of Winnctka, Illinois. Washburnc was described as one who "understood the value of t3cts and figures and fkncwl that they had to he collected by experts if they were to he credible" (Travcrs, 1983, p. 5l7).'
[)uring the 930s, as part of the progressive ducation movement, school discricts experimented with curricula based on the writings of John l)cwey. As

such curricula were developed and tried, they were evaluated, albeit informally. Critics of progressive high school curricula were unimpressed with these evaluations, however, and maintained that students educated in progressive high school curricula would rare poorly in higher education programs when compared to students educated in conventional Carnegie—unit curricula. This controversy lcd to the landmark Eight Year Study, which included a formal plan of evaluation that remains popular today. Ralph Tyler, employed in 1932 as director of the evaluation staff of the study, conceptualized the objectives—based approach to educational evaluation and developed instruments and procedures to measure a wide range of educational outcomes. In reporting their work on the Eight Year Study, Smith and Tyler (1942) provided an evaluation manual that was to dominate thinking in educational evaluatioi for the next quarter century. Even today, evaluators who employ objectives as the basis for determining whether a curriculum or program is a success (that is, if the objectives are achieved, the curriculum is judged to be successful) are still often referred to as Tvlerian evaluators. Later work by Bloom and others (1956), Taba (1962), Krathwohl, Bloom, and Masia (1964), Metfessel and Michael (1967), Bloom, Hastings, and Madaus (1971), and many others followed in the footsteps formed in the 1930s by Ralph Tyler. The United States National Assessment of Educational Progress (NAEP) was also conceptualized by Tyler in the 1960s, following the approach used in the Eight Year Study. Many current statewide testing programs in the United States are also based on the NAE1 design.

The 13Os also witnessed a growing influence among national and regional school accreditation agencies in the United States. Although they had been on the scene since the late 1800s and early 1900s, these agencies did not reach their pinnacle of power until the 1930s, for there were charters, standards, and memberships to be developed. Accreditation replaced the system of school inspections. based on a Western European school evaluation approach, thatwere common in many states in the 1890s (Glass, 1969). Using *Evaluatiue Criteria,* published by the National Study of School Evaluation, teams of outside educators were sent to review self—study reports of member institutions and to make their own observations. Recommendations from the team determined member institutions' status.

Unlike the Eight Year Study that concentrated on the outcome'sof schooling, the accreditation movement concentrated on the resources and processes used in schools. For example, accrediting agencies developed guidelines and criteria to monitor the adequacy of facilities, qualifications of staff, and appropriateness of program design, rather than assessing the educational status of graduates. Accreditation has been a highly influential part of American evaluation since the late 1800s. With the establishment of formal accrediting agencies for schools and colleges came the institutonalization of at least a quasi—evaluation process in American education.

The 1940s and early 1950s generally saw a period Qf consolidation and application of earlier evaluation developments. School personnel devoted their energies to testing and test development, accreditation, school surveys, and the formation or selection 0f acceptable objectives for education. The 1950s and early 1960s ilso

saw considerable technical development, building on the Tylerian base. For example, taxonomies of possible educational objectives were published, beginning with the influential *Taxonomy of Educational Objectives: Handbook I: Co,gnitive Domain* (Bloom and others, 1956). "Bloom's taxonomy," as it came to be popularly called, defined in explicit detail a hierarchy of thinking skills applicable to various content areas. This document continues to be a standard tool both in testing and in curriculum development, design, and evaluation. A sequel companion volume, entitled *Taxonomy of Educational Objectives: Handbook II: Affective Domain* (Krathwohl and others, 1964) has become popularly known as "Krathwohi's taxonomy." It provided, in the same detail as Bloom's taxonomy, an organized structure for evaluating and teaching feelings, emotions, and values. As reference tools for the evaluator, these taxonomies have proven indispensable.

Prior to 1965, the most dramatic change in educational evaluation resulted from the Soviet Union's launch of Sputnik I in 1957. American reaction was nearly immediate. With passage of the National Defense Education Act of 1958, millions of dollars were poured into development of new educational programs. Major new curriculum development projects, especially in mathematics and science (biology, chemistry, and physics), were initiated across the country. Subsequently, funds were made available to evaluate these curriculum development efforts.

The relatively few evaluation studies that resulted revealed the conceptual and ethodological impoverishment of evaluation in that era. In nianv eases, the designs were inadequate, the data invalid, the analyses inaccurate, and the reports irrelevant to the important evaluation questions that should have been posed. Most of the studies depended on idiosyncratic combinations and applications of concepts and techniques from experimental design, psychometrics, curriculum development and, to a lesser extent, survey research. 'Theoretical work related to educational evaluation per se was almost nonexistent. Few scholars had yet turned their attention to developin& evaluation plans applicable to education. Thus, educational evaluators were left to glean what they could from other ficids. 1 hat their gleanings were meager was noted b Cronbach (l%3) in a seminal article criticizing evaluations of the past as largely unhelpful and calling for new dtrcctions. Cronhach believed educational evaluation should help developers improve their products during early stages, nor just appraise their efiectiveriess once they were in the marketplace. Although Cronbach's recoin inend,itions had little rn mediate impact, they did stimulate sufficient dialogue ;i niong evaluation specialists to launch a greatly expanded conception of evaluation, is will be discussed later.

About the time of Cronbach's pronouncements. civil rights and concern for the disadvantaged began to gain increasing attention at the federal level. The Civil Rights Act of 1964 led to the Coleman Study iii 1965— 1966 that focused on equality of opportunity. for niinority children. Eve more important to educational evaluation, however, was the passage of the Eleiiientar arid Secondary Eddicatlon Act (ESEA) of' 965. which authorized sevral educational research, development, and disseniination activities. Hut the largest single coniponent ***(H*** the bill was Title I (later Chapter 1) educational programs for disadu iitaged youth destined to be the most costly federal education program in .\nierican history.

As Congress began its deliberations on the proposed ESEA, ***it*** became apparent that if passed. this bill would result m telis ot thousands of federal grants to local eduation agencies, intermediate and state agencies, and universities. Concerns began to be expressed, especially on the Senate Hoor, that there was absolutely no assurance that the large amounts of money

authorized *for* education would he spent as intended. ESEA was by fr the most comprehensive and ambitious education bill ever envisioned, and it was noted that education did not have an impressive record of providing evidence that federal monies thus expended resulted in any real educational improvements. Indeed, there were some in Congress who felt federal funds allocated to education prior to ESEA had sunk like stones into the morass of educational programs with scarcely an observable ripple to mark their passage.

Robert F. Kennedy was among those senators who forcefully insisted ESEA carrY a proviso requiring educators to be accountable for the federal monies they received. The were to tile an evaluation report for each grant showing what effects had resulted from the expenditure of the federal funds. Although only partially successful (the final version of the bill required evaluation reports under only two of its Titles—I and III), these efforts led to the first major federal mandate for educational evaluation. Translated into operational terms, this meant that thousands of educators were for the first time rcquiri'c l to spend their time evaluating their own efforts. Project evaluations mandated by state and federal governments have since become standard practice. with evaluation emerging as a political tool to control the expenditure of public funds.

Educational Evaluation: 1965—Present

In 1965, American educators were unprepared to **respond** effectively to the new **evaluation mandate.** Few **had** any expertise in evaluation. As **a result,** many **school** districts released their best teachers from classroom duties and pressed them into service as Title I or Title III evaluators. Their only qualifications for the job were experience and training as teachers—hardly relevant credentials for the position at hand. Even those who. possessed technical expertise were ill—prepared for the new demands of the federal mandate; their training in experimental research design, measuremen and statistics—while relevant—did not prepare them adequately to conduct evaluation. Nevertheless, some help is better than none, and these technical experts were widely employed. Using the resources known to them, they borrowed heavily from the behavioral, social, and educational sciences to conduct educational evaluation with, as we shall see, dubious results.

That many of the resulting "evaluations' would be inadequate was inevitable. Egon Guha was one who, after an analysis of the evaluation plans contained in a sample of Title Ill project proposals, concluded that

It is very dubious whether the results of these evaluations will be of much use to anyone. . None of these prod,uct evaluations will give the Federal Government the data it needs to review the general Tkle III program and to decide how the program might be reshaped to be more effective. (Cuba, 1967. p. 312)

Lack of trained personnel was not the only reason for the poor response to the ESEA evaluation mandate. In translating the legislation into operational terms, the United States **Office of** Education (USOE) had failed to provide adequate guidelines for the local evaluator. In the absence of such guidelines, evaluation designs for each project had to be created *de' tivo* by inexperienced personnel.

It seems likely that this failure to provide useful guidelines resulted more from lack of knowledge about what a good evaluation should include than from lack of effort on the part of USOE personnel. The expertise and methodology needed was either not **available or else** it was not adequate to address new evaluation needs. Few scholars had concerned themselves with generalizable evaluation plans for use by local.etaluators Theoretical work in evaluation was almost nonexistent.

The resulting vacuum was quickly filled, however, during the period of 1967— 1973, as academics developed new approaches, strategies, and methods for evaluators to use in these federal projects. Beginning in 1967, some observers began circulating their notions about how one should conduct educational evaluations. Their efforts produced several new evaluation "models" touted by the authors as responsive to the needs of Title I and Title III evaluators, and relevant to the ongoing curriculum development efforts sparked by Sputnik. New evaluation approaches were also proposed by educationists in England, Australia. Israel, Sweden, and other countries.

Collectively, these new conceptualizations of evaluation greatly broadened earlier vicws. As these frameworks for planning evaluation studies were refined. evaluators began increasingly to rely on them for guidance. Although these models couldn't begin to solve all the evaluation problems of local evaluators, they did help them circumvent several of the more treacherous pitfalls common to earlier evaluation studies. Problems caused by mindless application of objectives-based (Tylcrian) evaluation methods to every evaluation were revealed. The need to evaluate unintended outcomes of a curriculum was recognized. Values and standards were emphasized, and the importance of making judgments about merit and worth was made clear. These new and controversial ideas spawned dialogue and debate that **fed a** developing evaluation vocabulary and literature. The result has beer) a plethora of evaluation articles and books in the past two decades, containing at least 4() formalized or semiformalized ecaluation "models" proposed for use in education. (Fortunately, these models can he organized into several more generalizable evaluation approaches, as shown later in Part Two.)

I)uring this period, when the ESEA gave such profound impetus to educational evaluation, other developing trends increased the emphasis on evaluative processes. The gro''th of teacher militancy, union demands, and calls for civil rights reforms in education all required refined capabilities in evaluation. Thc increasing public outcry for educational accountability caused educators to rethink education's responsibilities and outcomes and ways of documenting them. Concerns over educational achievement led to student testing at federal and state levels. The National Assessment of Educational Progress, begun ia 1964 under the direction of Ralph Tyler, continues today with an annual assessrnent(of performance based on a national sample of students. It was not long until stat} departments of education began designing state assessment systems, and state legislatures began requiring

school reports on student achievement in subjects like reading and mathematics. Most states today conduct some type of statewide testing pragralil.

During the late 1960s and early 1970s, professional associations began encouraging their members to grant evaluation more serious attention. For example, the American Educational Research Association initiated a monograph series in curriculum evaluation. The Association for Supervision and Curriculum Development (ASCD) published evaluation monographs that encouraged curriculum developers to employ better evaluation techniques in assessing the worth of their products. More important, the 1970s gave rise to new professional asso— ciations for evaluators. Although Division H in the American Educational Research Associatiol? had been created as a professional home for schoo!—based evaluators, no association had yet been created to serve the evaluation specialist exclusively. In 1975, Phi Delta Kappa International provided seed money to establish the Evaluation Network, an interdisciplinary professional association of evaluators. The Evaluation Network quickly grew to several thousand members and sponsored a quarterly publication,' *Evaluation News*. The Evaluation Research Society, conceived by Marcia Guttentag and established in 1976. also developed as a multidisciplinary professional association for evaluators. This society also sponsored several publications. Beginning in 1986, a merger of these two associations resulted in a new, broader based American Evaluation Association.

The United States government has made concerted efforts to support and improve evaluation of the nation's educational system. In 1967, the federal government created the Center for the Study of Evaluation, a federally supported research and development center at the University of California, Los Angeles. In 1972, the government created the National Institute of Education (NIE). The NIE focused one of its research programs on evaluation in education, supporting field research that added to our knowledge of evaluation methodology, and also funded research to adapt methods and techniques from other disciplines for use in educational evaluation. During this same period (1968 to 1977), the budget for the Office of Planning, Budgeting, and Evaluation in the United States Office of Education was reported to have grown 1,650 percent (McLaughlin, 1980). One of the most enduring evaluation efforts supported by this federal agency was the operation of a series of Technical Assistance Centers for Title I/Chapter I Evaluation. Initiated in 1976, this nationwide network of centers has gradually expanded its perspective from mandated Title I evaluation and reporting to a broad range of evaluation—related assistance services addressing nearly all aspects of Chapter (Title) I programs.

The professional literature in evaluation has grown exponentially during the past 20 years, with the appearance of (1) numerous manuals, anthologies, and textbooks on selected evaluation issues; (2) journals such as *Evaluation, Evaluation and Program Planning, Evaluation News, Educational Evaluation and Policy Analysis, Evaluation Quarterly, New Directions for Program Evaluation,* and *Evaluation Rc'view;* and (3) annual compilations of evaluation literature published in the *Evaluation Studies Review 4nnuals.* Talmage-(1982) reports that the evaluation literature began to burgeon rapidly around 1974.

A joint Committee on Standards for Educational Evaluation, created in the

United States in 1975, now includes representatives from most major professional educational associations in the nation.7 In 1981, this Joint Committee developed the *Standards for Evaluations of Educational Programs, Projects, and Materials,* the first organized statement of principles for sound educational evaluation. A parallel effort by the Evaluation Research Society in 1982 resulted in a second set of standards, proposed to guide program evaluation practices in the diverse fields represented by the Society's membership.

The past decade in educational evaluation may be called one of "professionalization," as the shared knowledge and experience of a great many evaluators in education grew and matured. As was noted earlier in this chapter, "...evaluation is making the transition fiom late adolescence to adulthood." Yet, it must continue to grow and adapt to changing conditions and demands. For example, economic austerity and political conservatism in many nations during the early l980s have led to retrenchment and reduction in many social and educational programs. especially those sponsored by national governments. The resulting decrease in demand for evaluation of large—scale, federally supported educational programs has led some commentators to make gloomy predictions about the future of evaluation in education and related areas. Some insightful analysts, however, have forecasted that evaluation of educational programs will continue, although the locus and tocus will shift increasingly to local education agencies as evaluation's role in improving educational programs becomes institutionalized.

APPLICATION EXERCISES

1. Investigate parallels between the growth of evaluation thought in education and the growth of thought in other educational functions such as curriculum development, testing, and school administration.

2. List several reasons why evaluation in American education prob.hlv vould not have developed as it did ii the federal governilient had not been involved in educational reform.

SUGGESTED READINGS

MADAUS, (., **AJRASIAN,** P.. **&** **KELLACI1AN,** T. (198(uj. Sc/wi'! *f1'cil'cni3s*. New York: McGraw—Hill.

tV1ADAUS, G. **STUFFLEBEAM.** I)., & **SCRIVEN.** M. (1983). Program evaluation: A historical overview. **III (;. MADAUS.** M. SCRIVEN, **&** 1). Srui+l.i:nLM (Eds.), Li'aluar,ii *ni,'cliIs. 1''icu'points* **oit** *educarn,nal and hwnan* seri'ices *ea'aiiiation* . Boston: K luwer—NijhotI.

IALMA(:E, H. (1982). Evaluation of programs. In H. E. MirEl. (Ed.). *Inc'clcipcdia* Of *tional rcscnrcfi* (5th ed). New York: The Free Press.

FRAVERS. *R.* (1983). *How rescarth has chanç'i'd .4nrerica,, cli0ols.* Kalaniizoo. Ml: Mthos Press.

Chapter 3
The Concept of Evaluation: An Overview

Orienting Questions

1. If a person said she was doing an evaluation, what mental images would you have of her work?
2. What are some alternative ways that evaluation has been defined? Which definition do the authors prefer, and why?
3. What are some similarities and differences between educational research and evaluation?
4. When a person measures educational outcomes, has she evaluated the program that produces those outcomes?
5. What are the major differences between *formative* and *summative* evaluations?
6. What difference does it make whether a formative or summative evaluation is conducted by internal or external evaluators?

In the previous chapter on the history of evaluation, the perceptive reader will have noticed that the term *evaluation* was used broadly to encompass many diverse activities, ranging from testing student achievement to conducting accreditation site visits. Also, several evaluation terms and concepts were used without definition, except those that were implicit in context. Although such a general level of discourse was appropriate for that chapter, it is necessary now to become more specific about basic concepts and distinctions in order to provide common concepts and vocabulary for later chapters.

Like many disciplines, educational evaluation has developed its own jargon. For example, "evaluand" is often used to refer to whatever is being evaluated, unless it is a person, who is then an "evaluee" (Scriven, 1981). Moreover, the same terms are often used by different writers to refer to very different concepts or activities; even the term *evaluation* has been used to refer to so many disparate phenomena that the result is a confusing tangle of semantic underbrush through which the student of evaluation is forced to struggle.

The purpose of this chapter is to define evaluation more precisely, differentiate it

from other related but different activities, and examine two basic distinctions important to educational evaluation. The remainder of this chapter is divided into five major sections: (1) definition of evaluation and several other related activities; (2) a discussion of evaluation as a form of disciplined inquiry; (3) a discussion of similarities in and differences between educational research and evaluation; (4) an effort to expand the definition of evaluation; and (5) an examination of two basic distinctions in evaluation.

DEFINITION OF EVALUATION AND RELATED ACTIVITIES

There is no widely agreed-upon definition of educational evaluation. Some educators equate evaluation with measurement. Others define evaluation as the assessment of the extent to which specific objectives have been attained. For some, evaluation is synonymous with and encompasses nothing more than professional judgment. Some view evaluation as primarily scientific inquiry, whereas others argue that it is essentially a political activity. There are those who define evaluation as the act of collecting and providing information to enable decision—makers to function more intelligently. And so on.

In Chapter 4, we will discuss at greater length alternative views of evaluation and how these differing conceptions lead to widely varied types of evaluation studies. Our purpose here is only to sort out evaluation from among other educational activities and functions with which it often becomes confused.

Simple Verbal Definitions

Simple verbal definitions8 of research, evaluation, and measurement are never fully satisfactory. Research, evaluation, and measurement are complicated endeavors; they are ultimately no more than hypothetical constructs that allow us to speak with consistency about certain approaches to the production of information or knowledge. The meaning of the words can be seen by examining the ways scholars use the terms *research, evaluation,* and *,neasure,nent* both in their writing and conversation.

This is not to argue that no attempt should be made to differentiate these activities on an abstract, verbal level. Indeed, confusion among the terms accounts for much wasted motion in educational scholarship and complicates the already difficult jobs of conducting evaluations or teaching others how to conduct evaluations.

Despite the shortcomings of simple verbal definitions, they can serve as a point of departure, providing necessary precursors for later discussions.

Evaluation is the determination of a thing's value. ' In education, it is the formal determination of the quality, effectiveness, or value of a program, product, project, process, objective, or curriculum. Evaluation uses inquiry and judgment methods, including: (1) determining standards for judging quality and deciding whether those standards should be iclativc or absolute; (2) collecting relevant

information; and (3) applying the standards to determine quality. Evaluation can apply to either current or proposed enterprises.

Research is systematic inquiry aimed at obtaining generalizable knowledge by testing claims about the relationships among variables, or by describing generalizable phenomena. The resulting knowledge, which may generate theoretical models, functional relationships, or descriptions, may be obtained by empirical or other systematic methods and may or may not have immediate application.

Measurement (which is usually conceived of more broadly than mere testing) is the quantitative description of behavior, things, or events (mainly how much of a quality, or characteristic, an individual thing or event possesses). Measurement is simply a process for collecting the data on which research generalizations or evaluative judgments will be made. It is a key tool in evaluation and research; engaging in measurement is not, in and of itself, either research or evaluation.

Research is clearly an enormously complex undertaking that goes far beyond simple evaluative findings (for example, Program A is better than Program B on a particular criterion) by trying to ascertain the causes for those findings. The complexity of unraveling causes makes research a luxury few school districts can afford. Research has many of the trappings of evaluation and shares with it many common activities, but **it** lacks evaluations explicit judgments of quality.

*Evaluation **research,*** a term much used in the social sciences and sometimes in education as well, is, in our opinion, something of a misnomer, and its useseems to obscure more than clarify. Rather than try to define **it** here, we shall explain in a subsequent section why we avoid using this term in our treatment of educational evaluation.

Why Definition **Is Important**

Some readers may think that entirely too much fuss is being made over defining "evaluation." Meanings of words are critical, however, because words influence action.

Evaluation is complex. It is not a simple matter of stating behavioral objectives, building a test, or analyzing data, though **it** may include these activities. A thorough evaluation contains elements of a dozen or more distinct activities, the precise combination influenced by time, money, expertise, the good will of school practitioners, and many other factors. But equally important (and more readily influenced) is the image the evaluator holds of evaluation work: its responsibilities, duties, uniqueness, and similarities to related endeavors.

We frequently meet persons responsible for evaluation whose sincere efforts are negated by the particular semantic problem addressed in this chapter. By happenstance, habit, or methodological bias, they may label the trial and investigation of a new curriculum "research" or "experiment" instead of "evaluation." The inquiry they conduct will be different because they have chosen to call it a "research project" or an "experiment" rather than an "evaluation." Their choice influences the literature they read (it will deal with research or experimental design), the consultants they call in (only acknowledged experts in designing and analyzing

experiments), the way they report the results (always in the best tradition of the *American Educational Research Journal* or the *Journal of Experimental Psychology),* and the criteria used to judge how, well the study was done (emphasizing objectivity, internal and external validity, and other standards used to judge research practices). These are not necessarily the paths to relevant data gathering or rational decision making about curricula. By way of analogy, established canons of practice exist that underlie sound medical research procedures, but those are not the most relevant procedures that the surgeon must employ when performing a tonsillectomy.

Important as it is to define evaluation, therefore, a final, authoritative definition is illusive, if not impossible, because of the lack, of consensus about this phenomenon and just what it comprises. Talmage (1982) has made this point well, noting that, "Three purposes appear most frequently in definitions of evaluation:
(1) to render judgments on the worth of a program; (2) to assist decision-makers responsible for deciding policy; and (3) to serve a political function" (Talmage, 1982, P. 594). Talmage also notes that, while these purposes are not mutually exclusive, they receive different emphases in different evaluation studies, thus negating the possibility of any single, correct definition of evaluation.

In general, we have no quarrel with Talmage's analysis, but we would note one point of departure. For us, the first purpose she lists for evaluation—to render judgments of the value of a program—is not just a purpose for evaluation but *is* evaluation. Conversely, the other purposes do not describe what evaluation is but rather what it is *used for*. We choose, therefore, to define evaluation as the act of rendering judgments to determine value—worth and merit—without questioning or diminishing the important roles evaluation plays in decision-making and political activities. We agree with those who note that the important issues in education usually focus on decisions to be made, not judging value for the sake of judging value. But while this ties our definition of evaluation closely to the decision—making context in which evaluation is typically used, it does not narrow the definition of evaluation to serve only decision making.

To understand our view of evaluation better, it is necessary to consider a broader issue, namely, the role of disciplined inquiry in reshaping and revitalizing educational programs and practices.

EVALUATION AS DISCIPLINED INQUIRY

Both educational leaders and the public rightly expect the scientific method to play an important role in educational improvement. Cronbach and Suppes phrased this expectation well in their discussion of disciplined inquiry for education:

There has been agreement, both within and without the ranks of educators, that systematic investigation has much to offer. Indeed, there is agreement that *massive 1astint changes in education cannot safely be made except on the basis of deep oblective inquiry*.

(Cronbach & Suppes, 1969, p. 12)

Such systematic investigation, termed' by Cronbach and Suppes as 'disciplined inquiry," can take many forms (for example, a laboratory experiment or a mail

survey); however, there is a quality that is common to cach. As Cronbach and Suppes put it,

Disciplined inquiry has a quality that distinguishes it from other sources of,ppinion and belief. The disciplined inquiry is conducted and reported in such a way that the argument ***can*** be painstakingly examined. The report does not depend for its appeal on the eloquence of the writer or on any surfijce plausibility. The argument is not justified by anecdotes or casually assembled fragments of evidence. Scholars in each field have developed traditional questions that serve as touchstones to separate sound argument from incomplete or questionable argument. Among other things, the mathematician asks about axioms, the historian about the authenticity of documents, the experimental scientist about veritiability of observations. Whatever the character of a study, if it is disciplined the investigator has anticipated the traditioml questions that are pertinent. He institutes controls at each step of information collection and reasoning to avoid the **sources of error** to which these questions refer. If the errors cannot be eliminated, he takes them into account by discussing the margin for error in his conclusions. Thus the report of a disciplined inquiry has a texture that displays the raw materials entering the argument and the logical processes by which they were compressed and rearranged to make the conclusion credible.

Disciplined inquiry does not necessarily follow well—established, formal procedures. Some of the most excellent inquiry is free—ranging and speculative in its initial stages, crying what might seem to be bizarre combinations of ideas and procedures, or restlessly casting about for ideas. .. . **But..** . fundamental to disciplined inquiry is its central attitude, which places a premium on objectivity and evidential test. (Cronbach & Suppes, 1969, pp. 15—16, 18)

Inquiry, thus defined, encompasses several common activities in education, among them educational evaluation. Of course, systematic inquiry is not the only hallmark of evaluation. As Cronbach has also noted, evaluation is both a political and scientific activity:

Evaluation does, of course, draw on scientific tradition. It has to be judged in part by scientific ideals, and it surely should use all the techniques and principles from relevant science that it can. But science is only part of the story, and, I would say, a subordinate part. If evaluation is not primarily a scientific activity, what **is** it? It is first and foremost a political activity, a function performed within a social system. (Cronbach, 1977, p. 1)

Although, as noted earlier, one might better view political activity as a major ***use*** of evaluation, rather than a description of what evaluation is, Cronbach's point should not be lost. Evaluation is not only a scientific activity, as we shall discuss in some detail later in this text. Yet, evaluation ***is*** based on certain ideals and knowledge, and it differs from many forms of social or educational service in its insistence upon systematic, disciplined inquiry in the broad sense in which this concept was originally defined by Cronbach and Suppes.

Research is obviously an additional example of disciplined inquiry. As noted earlier, both educational research and evaluation have a great deal in common; both depend very heavily on empirical inquiry methods and techniques. If one were to watch tests being administered to students, or observe recording time—on- task data in classrooms, it would often ***be*** difficult to know whether the ongoing activity was research or evaluation. When one considers the purpose for which the

activity is conducted, research and evaluation have less in common, because the objective of the inquiry is quite different for these activities (as will be shown later in this chapter).

Despite differences in purpose, educational evaluation and research are frequently mistaken for one another. This may not seem particularly serious to the casual observer, but such confusion can have serious consequences indeed for the *researcher* or *evaluator*. The best efforts of an investigator whose responsibility is evaluation but whose conceptual perspective is research usually prove inadequate as either research or evaluation. Because it is especially crucial for the evaluator to distinguish research from evaluation, we will attempt to sharpen the distinction in the next chapter.

SIMILARITIES AND DIFFERENCES IN EDUCATIONAL EVALUATION AND RESEARCH

Thus far we have only attempted to differentiate between broad conceptions of research and evaluation. Distinguishing between such activities is at best extremely difficult and is made more so by the fact that each can be subdivided into several distinct activities. Many researchers have proposed their favored schemes for classifying research (Borg & Gall, 1983; Kerlinger, 1975). Similarly, several evaluators have attempted to classify varying types of evaluation (Worthen & Sanders, 1973; Stufflebeam & Webster, 1980; Brinkerhoff, Brethower, Hluchyj. **&** Nowakowski, 1983; Worthen, 1984). Although important differences exist among these various types of research and evaluation, there is an ingredient common to all: production of knowledge, however general or specific, not previously available. With this commonality in mind, we will now turn our attention to the more important differences among various types of research.

The distinction between basic and applied research seems well entrenched in educational parlance. Although these constructs might more properly be thought et as the ends of a continuum rather than as a dichotomy, they are useful for differentiating between broad classes of activities. **Ii** The distinction between the two also helps in subsequent consideration of how research relates to evaluation. l)efinitions of basic and applied research provided nearly three decades ago by the National Science Foundation still serve to make the distinction:

Basic research is directed toward increase of knowledge; it is research where the primary aim of the investigator is a fuller understanding of the subject under study rather than a practical application thereof. *Applied* ***research*** is directed toward pr&ctical applications of knowledge. (National Science Foundation, 1960. p. 5)

When successful, applied research results in plans or directives for development; basic research does not. In applied research, the knowledge produced must have almost immediate utility, whereas no such constraint is imposed on basic research. Basic research is intended to enhance understanding, and practical utility of the knowledge is not an issue.

Two variants of applied research are "institutional research" and "operations

research," activities imed at supplying institutions or social systems with relevant data To the extent that the conclusions resulting from inquiriesof this type are generalizable, at least across time, these activities may appropriately be labeled ***research.*** However, where the object is nongeneralizable information on performance of a program or process, the label *evaluation* might be more appropriate.

Action research, a concept that became popular and then all but disappeared during the 1950s, was also seen as a type of applied research. Proposed as a way to integrate studies into practical educational settings, the idea of action research was that researchable problems or issues are best identified in the classroom or school— where the action is. Immediate research work could be used to generate empiri—. cally supported ans'vers that would then be applied. Although the concept of action research had much appeal, it proved to be unworkable for several reasons: Educators could not be released for enough time to carry **Out** needed research, insufficient numbers of educators were trained in research methods, the cost of doing the research did not justify the benefits that were derived, and the problems that were able to be studied in a short period of time were trivial. The important problems required time, personnel, and funds that were unavailable to educational agencies (Hodgkinson, 1957; Clifford, 1973).

Evaluation has sometimes been considered a form of applied research that focuses only on one social program, one curriculum, or one classroom lesson. This view ignores an obvious difference between the two: the level of generality in the knowledge produced. Applied research (as opposed to basifl research) is aimed at producing generalizable knowledge relevant to providing a solution to a *general* problem. Evaluation focuses on collecting *specific* information relevant to a parti cula problem, program, or product. -

This brings us again to the term *evaluation research.*

Evaluation Research Defined and Discarded

The term ***evaluation*** *research* was popularized in the early 1970s, beginning with Caro's 1971 book, *Readings in Evaluation Research.* Used chiefly among social scientists, the term was adoptcd by the Evaluation Research Society and has become the model terminology used to describe evaluation in that association and in many evaluation publications by social scientists.

Those who find meaning in the terms *evaluation research* or *evaluative research* (which we shall hereafter treat as synonymous) are quick to separate evaluation research from evaluation. For example. Rossi states that

evaluation research is not equivalent to evaluation. To the extent that an evaluation is based on empirical evidence collected in ways that are susceptible to replication and treated with due regard to issues of internal, external, and construct validity, then the evaluation in question is evaluation research. The disciplines that have concerned themselves with the methodological problems of empirical research on social systems have been the social sciences *(from* which I would not exclude parts of education). Whether or not a given individual or organization engaging in evaluation, research regards

her/himself or itself as doing social science, he/she/it is nevertheless doing so. The

evaluation research in question may be conducted very well or very poorly, but it is social science research.

This is not to deny the possible validity of a connoisseurial approach to evaluation, in which expert judgments based on eccentric (but insightful) observations play major roles. Indeed, such evaluations may be worth more per resources expended. But they are not

social science nor are they evaluation research. (Rossi, 1982a, p. 61)

Rossi views evaluation research, therefore, as the application of social science research methods to evaluation issues. Talmage seems both to support and broaden this perspective when, in attempting to differentiate *evaluation research* from *research* and *experimental research,* she says:

Evaluation research studies usually lack replicabiliry because the system, program. or phenomenon being studied is dynamic; that is, it is in operation, changeable, and taking place in a naturalistic or field setting. Whereas the canons of scientific rigor are applied to evaluation research as far as possible, it is necessary to augment the study with

descriptions of contextual variables, and to utilize the methodologies and perspectives of various disciplines in order best to understand the processes and functioning of the system, program, or phenomenon under study.

Evaluation of educational programs is one area of evaluation under the rubric of evaluation research; it. . . applies the full range of evaluation research activities from initial design and planning through program implementation and institutionalization. (Talinage. 1982, pp. 594—595)

We have no particular argument with Rossi or Talmage on matters of substance. but we do disagree on interpretation. We agree that research is not equivalent to evaluation, regardless of its modifying adjectives, for, as argued earlier, research and evaluation differ in purpose even when they use the same methods and techniques. But we disagree that an activity should be termed evaluation *research* simply because it uses the experimental paradigm so favored among those social scientists who have shaped the form and direction of most social programs. Experimental research methods or a connoisseurial approach can both be used to evaluate—in other words, judge the worth or merit of—a social or educational program. It is not the tool that determines whether an activity is research or evaluation, but the purpose for which it is employed. It is not the use of social science methods that determines whether evaluation or research is being done, but the purpose for which the activity is being conducted. The neurosurgeon and the pathologist both use a scalpel, but to very diftèrcnt ends. Research and evaluation are no more synonymous, just because they may employ common approaches, tools, and techniques, than performing surgery is synonymous with doing an autopsy. Similarly, the form of inquiry we term *evaluation* may use research methodology (social science or otherwise) without being considered *research,* just as it may rely on cross—examination of expert witnesses to obtain data without being viewed as the practice of law.

l'ut simply, we see little to be gained by use of the term *ci'aliuuioti research.* And much clarity may be lost. To set what we hope will be a trend, there will be no further discussion or use of 'evaluation research" in this book.

Characteristics of Inquiry that Distinguish Evaluation from Research

It should be apparent from our previous discussion that some types of evaluation and research'2 resemble one another in certain ways. However, common characteristics shared by research and evaluation should not be allowed to obscure their fundamental differences. Contrasting the "purest" and most distinct forms of each activity admittedly results in oversimplifications, but should clarify some major points that may have been obscured in the previous discussion.'3

Before focusing on key differences between research and evaluation, we should reiterate an earlier caution from Stake and Denny (1969), which is still timely:

The distinction between research and evaluation can be overstated as well as understated. The principal difference is the degree to which the findings are generalizable beyond their application to a given product, program, or locale. Almost always the steps taken by the researcher to attain generalizability tend to make his inquiries artificial **or irrelevant** in the eyes of the practitioner. The evaluator sacrifices the opportunity to manipulate and control but gains relevance to the immediate situation. Researcher and evaluator work within the same inquiry paradigm but play different management roles and appeal to different audiences (Stake & Denny, 1969, p. **374)**

It is not our intent to overstate the distinction between research and evaluation. Yet, at a time when each is frequently mistaken for the other, **to** the detriment of both, we see a need for emphasizing differences more than similarities. What follows are 12 characteristics of inquiry that distinguish between 'pure" forms of research and evaluation.

Motivation of the Inquirer. Research and evaluation are generally undertaken for different reasons. Research satisfies curiosity by advancing knowledge; evaluation contributes to the solution of practical problems through judging the value of whatever is evaluated. The researcher is intrigued; the evaluator (or, at least, her client) is concerned.

The researcher may believe that her work has great long—range implications for problem solving, but that is not her primary motivation. If she is very nimble, she won't get bogged down in the seeming paradox that *policy* decisions supporting basic inquiry for its practical payoffs do *not* imply that researchers should focus on practical solutions.

Objective of the Inquiry. Research seeks *conclusions;* evaluation leads to *decisions.* Cronbach and Suppes (1969) distinguished between *decision-oriented* and *conclusion-oriented inquiry* this way:

In a decision—oriented study the investigator s asked to provide information wanted by a decision-maker: a school administrator, a government policymaker, the manager of a project to develop a ne'.v biology textbook, or the like. The decision-oriented study is a commissioned study. The decision-maker believes that he needs information to guide his actions and he poses the question to the investigator. The conclusion-oriented study, on the other hand, takes its direction from the investigator's commitments and hunches. The educational decision—maker can, at most, arouse the investigator's interest in a problem.

The latter formulates his own question, usually a general one rather than a question

about a particular institution. The aim is to conceptualize and understand the chosen phenomenon; a particular finding is only a means to that end. Therefore, he concentrates
on persons and settings that he expects to be enlightening. (Cronbach **&** Suppes, 1969, pp. 20—21)

Conclusion—oriented inquiry is here referred to as research; decision—oriented inquiry typifies evaluation as well as any three words can.

Laws vs. Descriptions. Briefly stated, research involves *nomothetic* (lawgiving) activities, and evaluation involves *idiographic* (descriptive of the particular) activities. Research is the quest for laws—that is, statements of relationships among two or more variables. Evaluation seeks to describe a particular thing and its unique context with respect to one or more scales of value.

Role of Explanation. Considerable confusion exists about the extent to which evaluators should explain ("understand") the phenomena they evaluate. We do not view explanations as the primary purpose of evaluation. A fully proper and useful evaluation can be conducted without explaining *wha caused* the product or program being evaluated to be good or bad or *hote'* it produces its effects. It is fortunate that this is so, for educational evaluation is so needed and credible explanations of educational phenomena are so rare, Of course, it would be wonderful if an evaluation of an outstanding training program managed also to explain what it was that made the program work so well, just so that good fortune did not lead to similar expectations for all educational evaluations.

Autonomy of the Inquiry. Science isan independent and autonomous enterprise. At the beginning of his classic. *The Conduct of Inquiry*. Kaplan (1964) wrote:

It is one of the themes of this book that the various sciences, taken together, are not colonies subject to the governance of logic, methodology, philosophy of science, or any other discipline whatever, but are, and of right ought to be, free and independent. Following John I)ewey. I shall refer to this declaration of scientific independence as the principle of *autonomy of inquiry*. It is the principle that the pursuit of truth is accountable to nothing and to no one not a part of that pursuit itself (Kaplan. 1964, p. 3)

Not surprisingly, autonomy of inquiry proves an important characteristic for typifying research and evaluation. As was seen incidentally in the quote from Cronbach and Suppes, evaluation is undertaken at the behest of a client, but the researcher sets her own task. As will be seen later, the autonomy that the researcher and the evaluator enjoy to diftiring degrees has implications for how they should be trained and how their respective Inquiries are pursued.

Properties of the Phenomena Assessed. Educational evaluation attempts to assess the *value* of a thing, whereas educational research attempts to generate scientific *knowlcd,51e*. Except that knowledge is highly valued and thus worthwhile, this distinction serves fairly well to discriminate research and evaluation. The distinction can be given added meaning if mine is taken as synonymous with *social utility* (which is presumed to increase with improved health, happiness, and life expectancy, and if *scientific kuowle dç'e* is identified with two of its properties: (1) empirical verifiability, and (2) logical consistency.

Evaluation seeks to assess social utility directly. Research may yield indirect

evidence of social utility, insofar as empirical verifiability of general phenomena and logical consistency may eventually be socially useful. Valuing is the *sine qua non* of evaluation. A touchstone for discriminating between an evaluator and a researcher is to ask whether the inquiry she is conducting would be regarded as a failure if it produced no data on the usefulness of the thing being studied. A researcher answering strictly as a researcher will probably say no.

Generalizability of the Phenomena Studied. Perhaps the highest correlate of the research-evaluation distinction is the generalizability of the phenomena being studied. Three aspects of generalizability can be identified: (1) generalizability across time (Will the phenomenon—perhaps a textbook or a self-concept—be of interest 50 years hence?); (2) generalizability across geography (Is the phenomenon of any interest to people in the next town, the next province, across the ocean?); and (3) applicability to a number of specific instances (Are there many specific examples of the phenomenon being studied or is this the only one?). These three qualities of the object of an educational inquiry can be used to classify different inquiry types, as in Figure 3.1.

Three types of inquiry are represented in Figure 3.1: (1) program evaluation—the evaluation of a complex of people, materials, and organization which make up a particular educational program; (2) product evaluation—the evaluation of a medium of schooling such as a book, a film, or a recorded tape; and (3) educational research.

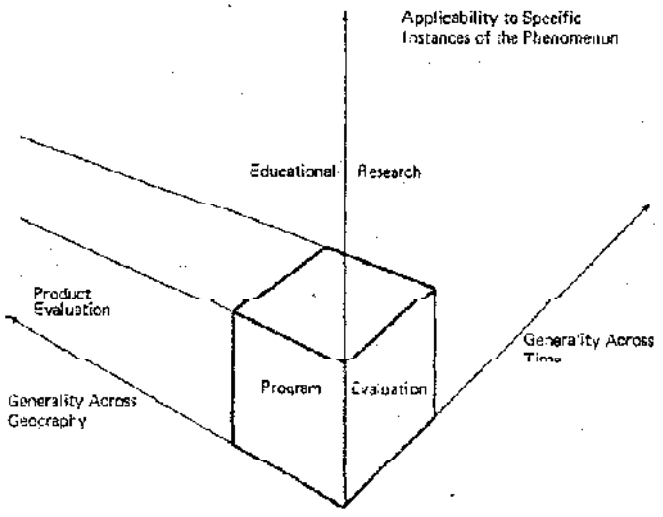


FIGURE 3.1 Three Inquiry Types Classified by the Generalizability of the Phenomenon Investigated*

*The property represented by each axis is absent where the three axes meet and increases as one moves out along each axis.

Program evaluation is depicted in Figure 3.1 as concerned with a phenomenon (an educational program) that has limited generalizability across time and geography. For example, the innovative "ecology curriculum" (including instructional materials, staff, students, and other courses in the school) in the Middletown Public Schools will probably not survive the decade, is of little interest to the schools in Norfolk that have a different set of environmental problems and instructional resources, and has little relationship to other curricula with other objectives. Product evaluation is concerned with assessing the worth of something like a new ecology textbook or an overhead projector that can be widely disseminated geographically, but that similarly may not be of interest 10 years hence, and that produces little or no reliable knowledge about education in general. Educational research focuses on concepts supposed to be re]aively permanent, applicable to schooling nearly everywhere, and relevant to numerous teaching and learning contexts.

Criteria for **Judging the** Activity. Two important criteria for judging the adequacy of research are internal validity (To what extent are the results of the study unequivocal and not confounded with extraneous or systematic error variance?) and external validity (To what extent can the results be generalized to other units—for example, subjects or classrooms—with characteristics similar to those used in the study?).

If one were forced to choose the most important criteria from among the several criteria that might be used for judging the adequacy of evaluation, the five most vital would probably be *accuracy* (the extent to which the information obtained is an accurate reflection—a one—to--one correspondence—with reality). *(rrdibility* (the extent to which the information is believable to clients who need it). *utility* (the extent to which the results are actually used), *feasibility* (the extent to which the evaluation is realistic, prudent, diplomatic, and frugal), and *propriety* (the extent to which the evaluation is done legally and ethically, protecting the rights of those involved.

Identifiable Clients. Research is often conducted with only the most nebulous idea of who may use the results. Conversely, evaluation is generally conducted for a well—defined audience or client group. Evaluators can identify such clients as policymakers, program managers, or concerned publics. Research results, on the other hand, are laid out for general perusal, by anyone who finds them interesting or useful.

Relevance of Time. Research sets its own time schedule—barring economic constraints. Evaluation, by contrast, is time—bound; both start—up and duration must adhere to strict schedules if the results are to be useful.

Disciplinary Base. Making educational research multidisciplinary may be a good suggestion for the research community as a whole, but it is doubtful that the individual researcher is well advised to attack her particular area of interest simultaneously from several differcnt disciplinary perspectives. Few persons can fruitfully work in the cracks between disciplines; most will find it challenge enough to deal with the problems of one discipline at a time.

That the educational researcher can afford to pursue inquiry within one para—

digm and the evaluator cannot is one of many consequences of the autonomy of inquiry. When one is free to define her own problems for solution (as the researcher is), she seldom asks a question that takes her outside of the discipline in which she was educated. Psychologists posc.questions that can be solved by the methods of their own disciplines, as do sociologists, economists, and other scientists. The seeds of the answer to a research question are planted along with the question. The curriculum evaluator enjoys less Ircedom in defining the questions she must answer. Hence, she *may* find **it** necessary to employ a wider range of inquiry perspectives and techniques in order to find her answers.

Preparation. The distinction just discussed, relating to disciplinary bases, has different implicarion for the education of researchers than for the education of evaluators. Research aimed at producing understanding of such phenomena as school learning, the social organization of schools, and human growth and development can best be accomplished from the perspectives afforded by the relevant social sciences. Consequently, the best preparation for many educational researchers is likely to be a thorough mastery of a relatively traditional social science discipline, coupled with application of the tools of that discipline to educational problems. Given the ambitiousness of preparing a researcher thoroughly in the methods and techniques of her chosen discipline, there is little point in arguing that any one researcher should receive training in more than one discipline. A graduate program that attempts to indoctrinate the same student in social psychology, physiological psychology, and microeconomic theory is likely to produce a researcher so superficially prepared in several areas that she can contribute in none.

By contrast, to the extent that the preparation of evaluators toucheson the traditional disciplines at all, it is best that several disciplines be sampled. Only through an interdisciplinary education can the evaluator become sensitive to the wide range of phenomena to which she must attend if she is to properly assess the worth of an educational program. That superficial exposure does not qualify her to conduct in-depth research is simply irrelevant.

The preparation of educational researchers need not be as broad in techniques of inquiry as that for evaluators. Most sociologists' work is none the worse for their ignorance of confounding relations in fractional factorial designs, and who would argue that an experimental psychologist suffers from ignorance of scalogram analysis? The education of researchers must be more concerned with substantive matters (such as Keynesian economics. Skinnerian operant conditioning, or the cognitive dissonance theory) than with methods and techniques. The evaluator, conversely, must be broadly familiar with a wide variety of methods and techniques because one evaluation might require her to use psychometrics, another sociometrics, a third econometrics, and so on.

In preparing both researchers and evaluators, some provisions must be made for the acquisition of practical experience. Doubtless there is no better practical experience for the research trainee than apprenticeship to a competent, practicing researcher. Worthen arid Roaden (1975) found that apprenticeship to one researcher over an extended period of time was positively correlated with subsequent researci

productivity. A protracted apprenticeship to a single evaluator would be inappropriate, however, because breadth of experience in a variety of settings is essential. Evaluation is unashamedly practical. Whereas the researcher can afford **to** ignore— indeed, *must* ignore—the countless practical constraints of contemporary schools in constructing her elegant idealizations, evaluation that ignores practicaliries is just bad evaluation.

TWO BASIC DISTINCTIONS IN EVALUATION

Prominent evaluation ,theorists differ widely in their views of what **evaluation** is and how it should be carried out. We will discuss these differences in **some detail** in Part Two. Despite these varying perspectives, however, some common concepts and distinctions exist about which there seems to be relatively little debate)6 These notions, though elemental, have proven powerful in shaping people's thinking about evaluation. In this section, we will discuss two basic distinctions in evaluation and how they apply to educational evaluation studies.

Formative and Summative Evaluation

Scriven (1967) first distinguished between the *[ornsative* and *sun,niath'e* **roles** of evaluation. Since then, the terms have become almost universally accepted in the field. Although in practice distinctions between these two types of evaluation may blur somewhat, it seems useful to summarize the major differences noted by Scriven, even at the risk of some oversimplification.

Formative evaluation is conducted during the operation of a program **to** provide program directors evaluative information useful in improving the program.'7 For example, during the development of a curriculum package, formative evaluation would involve content inspection by experts, pilot tests with small numbers of children, field tests with larger numbers of children and teachers in several schools. and so forth. Each step would result in immediate feedback to the developers, who would then use the information to make necessary revisions.

Summative evaluation is conducted at the end of a program to provide potential consumers with judgments about that program's worth or merit. For example, after the curriculum package is completely developed, a summative evaluation might be conducted to determine how effective the package is with **a** national sample of typical schools, teachers, and students at the level for which it was developed. The findings of the summative evaluation would then be made available to consumers.

Note that the audiences and uses for these two evaluation roles are very different. In formative evaluation, the audience is program personnel**—**in our example, those responsible for developing the curriculum. Summacive evaluation audiences include potential consumers (students, teachers, and other professionals). funding sources (taxpayers or funding agency), and supervisors and other officials, as well as program personnel. Formative evaluation leads to (or should lead to) decisions about program development (including modification, revision, and the

like). Summative evaluation leads to decisions concerning program continuation, termination, expansion, adoption, and so on.

It should be apparent that both formative and summative evaluation are essential because decisions are needed during the developmental stages of a program to improve and strengthen it, and again, when it has stabilized, to judge its final worth or determine its future. Unfortunately, far too many educators conduct only summative evaluation. This is unfortunate because the development process, without formative evaluation, is incomplete and inefficient. Consider the foolishness of developing a new aircraft design and submitting it to a "summative" test flight without first testing it in the "formative" wind tunnel. Educational test flights can be expensive too, especially when we haven't a clue about the probability of success.

Failure to use formative evaluation is myopic, for formative data collected early can help rechannel time, money, and all types of human and material resources into more productive directions. Evaluation conducted only when a project nears completion may simply come too late to be of much help.

Of course, the relative emphasis on formative and summative evaluation changes throughout the life of an educational program, as suggested in Figure 3.2, although this generalized concept obviously may not precisely fit any particular curriculum innovation.

Baker (1978) noted that two important factors which influence the usefulness of formative evaluation are *control* and *timing*. If suggestions for improvement are to be implemented, then it is important that the formative study collect data on variables over which program administrators have some control. Also, information that reaches administrators too late for use in improving the program is patently useless.

As will be noted later, many of the evaluation techniques and approaches described later in this book can be used as readily for formative as summative evaluation; the timing of their use and the purpose for which they are employed determines whether they play a summative or formative role.[18]

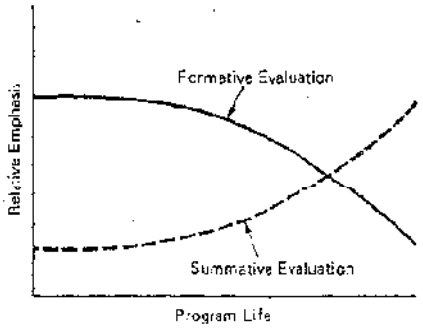


FIGURE 3.2 Relationship Between Formative and Summative Evaluation Across Life of a Curriculum Innovation

An effort to distinguish between formative and summative evaluation on several dimensions appears in Figure 3.3.[19]

Distinguishing Formative and Summative Evaluation in Practice. As with most conceptual distinctions, formative and summative evaluation are often not as easy to distinguish in practice as they seem in these pages. For example, if a program continues beyond a summative evaluation study, the results of that study may be used for both summative and, later, formative evaluation purposes. This may be one factor that led Stake (1969b) to suggest that formative and summative evaluation may be less distinguished by when they are conducted than by what "insiders" and "outsiders" want to know about educational programs.

> In evaluation circles, the terms *formative* and *summative* are heard more and more frequently. These terms have a dramatic effect, distinguishing between what is done *during development* and what is done *when development is finished*. For the purpose of choosing an evaluation strategy, I find this a trivial distinction. For most educational programs—correspondence courses or Montessori programs—development never ends. For a learner, there is a beginning and an end, but for the teacher, the programs are ongoing, ever evolving. What is important is that there are differences between what the "program people" want to know about their program and what "outsiders" want to know. We can make a non-trivial distinction between formative evaluation for the program developer who is planning ahead and trying to choose the best ingredients, and summative evaluation for anyone who is looking at the program, past or present, and who is trying to find out what it is and what it does. (Stake, 1969b, p. 40)

We accept the general wisdom of this caution and admit that sorting ongoing evaluations into formative and summative categories, as such, is unproductive. We

	Formative Evaluation	*Summative Evaluation*
Purpose	To improve program	To certify program utility
Audience	Program administrators and staff	Potential consumer or funding agency
Who Should Do It	Internal evaluator	External evaluator
Major Characteristic	Timely	Convincing
Measures	Often informal	Valid/reliable
Frequency of Data Collection	Frequent	Limited
Sample Size	Often small	Usually large
Questions Asked	What is working? What needs to be improved? How can it be improved?	What results occur? With whom? Under what condition? With what training? At what cost?
Design Constraints	What information is needed? When?	What claims do you wish to make?

FIGURE 3.3 Differences Between Formative and Summative Evaluation

maintain, however, that the distinction between formative and summarive cvalua—tion is relevant to a more important distinction between the purposes and audiences for these two evaluation roles.

Internal and External Evaluation

The adjectives *inernal* and *exerual* distinguish between evaluations conducted by program employees and those conducted by outsiders. An experimental year- round education program in the San Francisco public schools might be evaluated by a member of the school district staff (internal) or by a site—visit team appointed by the California State Board of Education (external). There arc obvious advantages and disadvantages connected with both of these roles.

The internal evaluator is almost certain to know more about the program than any outsider, but she may also be so close to the program that *she* is unable to be completely objective. Seldom is chcre as much reason to question the objectivity of the external evaluator (unless she is found to have a particular ax to grind) and this dispassionate perspective is perhaps her greatest asset. Conversely, it is difficult for an external evaluator to ever learn as much about the program as the insider knows. Note that when we say ***as** much*, we refer only to quantity, not quality. One often finds an internal evaluator who is full of unimportant details about the program but overlooks several critical variables. If these bits of key information are picked up by the external evaluator., as is sometimes the case, she may end up knowing much less *overall* about the project but knowing much more of importance. On the other hand, the internal evaluator is much more likely to be familiar with important contextual information (for example, the serious illness of the director's husband, which is adversely affecting the director's work) that would temper evaluation recommendations.

Anderson and Ball (1978) have noted that knowing whofunds an evaluation and to whom the evaluator reports largely determines the evaluator's financial and administrative dependence. They argue that independent relationships (that is, external evaluation) generally enhance the credibility of the study, a point with which we agree.

Others, however, question whether external evaluation really results in more credible evaluations. Cronbach and his colleagues (1980) are frankly unsympathetic to the view that the quality of evaluation is enhanced through the use of external reviewers. They take the position that credibility of evaluation studies lies in profession—wide arrangements that ensure the evaluator's freedom to be honest, not in the inherent objectivity of external review. And Campbell (1984) lists insistence on external evaluation as a major error that social science methodologists made in presenting their views of applied social science to government in the 1960s. He calls for cross—validation of studies akin to that conducted in the physical sciences as a better way to obtain objectivity than by depending on the "dogma" of evrernal evaluation.

Evaluators who have survived years of service as "in—house" evaluators or evaluation contractors would likely find these views somewhat quaint and far

removed from the political cauldrons in which they have seen (if not felt personally) employment locus and control of perquisites wreak havoc. Control of salary and perquisites can have a very real influence on objectivity and even truth, as Scriven (1976a) has noted in his examination of methods for bias control. Pronouncements to the contrary may have the ring of academic correctness but ignore the political practicalities. Similarly, these scholars' expressed lack of concern over evaluators who are too sympathetic to the program they are evaluating will be viewed as innocent by many veteran evaluators who have all too often seen colleagues' pro-program sympathies lead to unconscionable defenses of the status quo. The call for cross—yalidation of studies is compelling, using the referent of the physical sciences. But itioses its relevance and potency in the much less replicable realms of educational and social inquiry. It would be nice if educational evaluation studies could be cross—validated as readily as reading meters in the physical sciences. Alas, such is not the case. We continue to believe, despite the arguments of esteemed colleagues to the contrary, that the locus of the evaluator can have a profound effect on the results of an evaluation study. We will expand this point in the next section.

Possible **Role Combinations**

The dimensions of formative and summative evaluation can be combined with the dimensions of internal and external evaluation to form the two-by-two matrix shown in Figure 3.4.

The most common roles in evaluation might be indicated by cells 1 and 4 in the matrix. Formative evaluation is typically conducted by an internal evaluator, and there are clear merits in such an approach. Her knowledge of the program is of great value here, and possible lack of objectivity is not nearly the problem it would be in a summative evaluation. Summative evaluations are typically (and probably best) conducted by external evaluators. It is difficult, for example, to know how much credibility to accord a Ford Motor Company evaluation that concludes that a particular Ford automobile is far better than its competitors in the same price range. The credibility accorded to an internal summative evaluation (cell 3) of an educational program is likely to be no better. Summative evaluation is generally

FIGURE 3.4 Combination of Evaluation Roles
A

	Internal	*External*
Formative	1 Internal Formative	2 External Formative
Sutnrnatii'e	3 Internal Summative	4 External Sunimative

best conducted by an external evaluator or agency. In some instances, however. there is simply no possibility of obtaining such external help because of financial constraints or absence of competent personnel willing to do *the* job. In these cases, the summative evaluation is weakened by the lack of outside perspective, but it might be possible to retain a semblance of objectivity and credibility by choosing the internal summative evaluator from among those who are some distance removed from the actual development of the program or product being evaluated.

The difficulty in conducting objective, credible, internal summarive evaluations may be one reason for the report by Franks and Fortune (1984) that few local school districts in the United States arc administering any type of summative evaluations to prograths funded under Chapter 2 legislation. The view that relaxation of the federal evaluation mandate may be the real cause of this phenomenon is more cynical. Whatever the cause, it should be reiterated that both formative and summative studies of educational programs are needed for long—range improvement of our educational systems.

A very important role—that of the external formative evaluator shown in cell 2—is almost completely neglected in educational evaluations. As implied earlier, the internal evaluator may share many of the perspectives and blind spots of other program staff and, consequently, neglect to even entertain negative questions about the program. The external evaluator, who does not have lengthy familiarity with the program or its context, is much less likely to be influenced by a *priori* perceptions of its basic worth. This is not synonymous with saying that she is predisposed toward judging the program as ineffective. I-Icr orientation should be neither positive nor negative. It should be neutral, uninfluenced by close associations with either the program or its competitors. in essence, the external formative evaluator introduces a cold, hard look of reality into the evaluation relatively early—in a sense, a preview of what a summative evaluator might say. This fresh outside perspective is important and *can* preclude the disaster that frequently occurs when program staff self-consciously select criteria and variables they believe will prove their program successful, only to have an outside agency (a school board or site—visit team) later recommend terminating the program because their summative evaluation—which focused on other variables or criteria— suggested the program was ineffective. Wisdom would dictate the use of an outside evaluator as part of every formative evaluation. Scriven (1972), although accepting the utility of internal formative evaluation, also argued for this view, saying "it now seems to me that a producer or staff evaluator who wants good *formative* evaluation has got to use some external evaluators to get it" (Scriven, 1972, p. 2).

APPLICATION EXERCISES

1. List the types of evaluation studies that have been conducted in an educational institution of your acquaintance, noting whether the evaluator was internal or external to that institution. Determine whether each study was formative or

summative and whether the study would have been strengthened by having **it** conducted by someone with the *opposite* (internal/external) relationship to the institution.

2. Select one evaluation study and one research study with which you are familiar. Analyze both to see if they differ on the 12 dimensions identified in this chapter as useful in distinguishing between research and evaluation activities.

SUGGESTED READINGS

CRONBACH, L.J., & SUPPEs, P. (1969). *Research for tomorrou"s schools: Disciplined inquiry for education*. **New York: Macmillan.**

SCRIVEN, M. (1973). The methodology of evaluation. In B. R. **WORTHEN** & J. **R. SANDERs,** *Educational evaluation: Theory and practice*. Belmont, CA: Wadsworth.

STUFPLEBEAM, **D. L.,** & **WEBSTER,** W.J. (1980). An analysis of alternative approaches to evaluation. *Educational Evaluation and Policy Analysis,* 2(3), 5**—**19.

WORTHEN, B. R., & SANDERS, J. R. (1973). *Educational evaluation: Theory and practice*. Belmont, **CA: Wadsworth.**

elephant is like by piecing together reports of several blind people, each of whom happens to grasp a different portion of the elephant's anatomy. The evaluation literature is badly fragmented and is often aimed more at fellow evaluation theorists than at practitioners. Busy practitioners can hardly be faulted for not expending the time necessary to interpret and consolidate these disparate bits of knowledge.

We shall not solve the problem totally in this book. But this chapter and those that follow (Chapters 5 through 11) should clarify the varied alternatives useful in conducting evaluation. Perhaps by the end we will know enough about the elephant to recognize **it** when we see it.

DIVERSE CONCEPTIONS OF EDUCATIONAL EVALUATION

We have mentioned that at least three different views of educational evaluation have coexisted for the past 50 years. With the ascendency of the measurement movement, evaluation came to be defined as roughly synonymous with educational measurement. This orientation is evident today in the writing of such measurement specialists as Hopkins and Stanley (1981) and Sax (1980). Concurrently, formalization of school and university accreditation procedures led to the view that evaluation is synonymous with professional judgment. This view has persisted in many current evaluation practices where judgments are based on opinions of experts, whether or not the data and criteria used in reaching those judgments are clear. A third conception of evaluation emerged during Tyler's work on the Eight Year Study of the 1930s. In his work, evaluation came to be viewed as the process of comparing performance data with clearly specified objectives. This view is reflected in the next chapter.

Since 1965, new approaches have emerged as prominent methodologists and educationists turned their attention to evaluation methods. Many evaluation "models" emerged, ranging from comprehensive prescriptions to checklists of suggestions. In the absence of a good empirical base for determining the best way to evaluate educational programs, these models have greatly influenced present practices. Some authors opt for a systems approach, viewing evaluation as a process of identifying and collecting information to assist .decision—makers. Others focus on the importance of naturalistic inquiry, or urge that value pluralism be recognized, accommodated, and preserved and that those individuals involved with the entity being evaluated play the prime role in determining what direction the evaluation study takes. Some writers propose that evaluations be structured in keeping with legal or forensic paradigms so that planned opposition—both pro and con—is built in. And this barely dents the list of current alternatives.

As Guba and Lincoln (1981) noted, the idea that evaluation should determine the worth of an educational program is one of the few things these diverse approaches have in common. Talmage (1982) concurs, stattng that three purposes that appear most frequently in alternative evaluation proposals are "(1) to render judgment on

the worth of a program; (2) to assist decision makers; and (3) to serve a political function" (Talmage, 1982, p. 594).

On the other hand, the differing concepts and emphases in these variant evaluation approaches greatly influence the practice of evaluation. As noted previously, "the various models are built on differing—often conflicting—conceptions and definitions of evaluation, with the result that practitioners are led in very different directions, depending upon which model they follow" (Worthen, 1972b, p. 3). Let us consider an example.

If one viewed evaluation as essentially synonymous with professional judgment, the worth of the curriculum would be assessed by experts (as judged by the evaluation client) observing the curriculum in action, examining the curriculum materials, or in some' other way gleaning sufficient information to record their considered judgments. If evaluation is equated with measurement, the curriculum might well be judged on the basis of students' scores on standardized tests.

If evaluation is viewed as a comparison between performance indicators and objectives, behaviorally stated objectives would be established for the curriculum. and relevant student behaviors would be measured against this yardstick, using either standardized or evaluator—constructed instruments. (Note that in this process there is no assessment of the worth of the objectives themselves.)

Using a decision—oriented approach, the evaluator, working closely with the decision-maker, would collect sufficient information about the relative advantages and disadvantages of each decision alternative to judge which was best. However, although the decision-maker would judge the worth of each alternative, evaluation per se• would be a shared role..

If one accepted the authors' earlier definition of evaluation (see Chapter 3), the curriculum evaluator would first identify the curriculum goals and then, using input from appropriate reference groups, determine whether the goals were good for the students, parents, and community served. He would then collect evaluative information relevant to those goals as well as to identifiable side effects resulting from the curriculum. When the data were analyzed and interpreted, the evaluator would judge the worth of the curriculum and (usually) make a recommendation to the individual or group responsible for final decisions.

Obviously, the way in which one views evaluation has direct impact on the type of evaluation activities conducted.

ORIGINS OF ALTERNATIVE VIEWS OF EVALUATION

One might wonder why there are so many views of evaluation. But remember, there is no less disagreement about which learning theory is best, which research paradigm is preferable, or which school—management theory is most useful. It is no more reasonable to expect agreement on one approach to educational evaluation than it is to expect agreement on one way to teach or to run a school. Evaluation

scholars come *from varied* backgrounds and have different world views of education and inquiry. As needs for rethinking evaluation have appeared, writers— academics, for the most part—have been amazingly productive and creative in proposing responses to those needs. Often several theorists have worked concurrently and independently to address similar needs but have approached them in very different ways. Thus, the emerging views of evaluation have been widely divergent, sometimes based on conflicting assumptions, and often focused on differing goals or purposes. Evaluation literature will undoubtedly continue to reflect new views as other theorists share their insights.

Before presenting alternative evaluation approaches, we need to consider briefly the factors that led to differing views, including evaluators' diverse philosophical ideologies, cognitive styles, methodological preferences, values, and practical perspectives.

PHILOSOPHICAL AND IDEOLOGICAL DIFFERENCES

There is no univocal philosophy of evaluation, any more than there is a single. universally accepted philosophy of science. And perhaps that lack has not hurt us too much, for, as House (1983b) has noted, we have been doing without one for a long time. The lack of a guiding philosophy has not preempted extensive discourse and debate concerning philosophical assumptions about epistemology and value. Indeed, different approaches to establishing.truth or merit are largely responsible for the diversity of views about educational evaluation.

Objectivist and Subjectivist Epistemology

House (1980, 1983a, 1983b) has written extensively and thoughtfully about different philosophies of knowing or establishing truth (epistemology) and how they affect the approach to evaluation one would choose. He has grouped evaluation approaches into two categories: *objectivism* and *subjectivism.*

Objectivism requires that evaluation information be "scientifically objective": that is, that it use data collection and analysis techniques that yield results reproducible and verifiable by other reasonable and competent persons using the same techniques. In this sense, the evaluation procedures are "externalized," existing outside of the evaluator in clearly explicated form that is rcplicable by others and that will produce similar results from one evaluation to the next. Objcctivism is derived largely from the social science tradition of empiricism.

Subjectivism bases its validity claims on "an appeal to experience rather than to scientific method. Knowledge is conceived as being largely tacit rather than explicit" (House, 1980, ***p.*** 252). The validity of a subjectivist evaluation depends on the relevance of the evaluator's background and qualifications and the keenness of his perceptions. In this sense, the evaluation procedures are "internalized," existing largely within the evaluator in ways that are not explicitly understood or reproducible by others.

Objectivism has held sway in the social sciences and in educational inquiry for decades. There are, however, many criticisms of objectivist epistemology, as there are of logical positivism in educational science. Campbell (1984) states that, "Twenty years ago logical positivism dominated the philosophy of science.... Today the tide has completely turned among the theorists of science in philosophy, sociology, and elsewhere. Logical positivism is almost universally rejected" (Campbell, 1984, p. 27). Scriven (1984) argues that any lingering positivist bias in evaluation should be eliminated. And Guba and Lincoln (1981) have challenged the "infallibility" of the hypothetico-deductive inquiry paradigm because of its limitations in dealing with complex, interactive phenomena in dynamic, septic educational settings Although less sweeping and conclusive in his critique, House (1980) portrays objectivism as inattentive to its own credibility, presuming validity because of its methodology and, therefore, credible only to those who value such a methodology. He also notes that objectivism conceals hidden values and biases of which its adherents are unaware, because even the choice of data—collection techniques and instruments is not value—neutral, an assumption seemingly taken for granted by objectivist evaluators.

To counter the objectivist hold upon the methodologies of evaluation in education and related areas, criticisms of objectivism have been extreme. Yet sub— jectivism has been no less soundly criticized, especially by those who see its procedures as "unscientific" and, therefore, of dubious worth. Critics (for example, Boruch & Cordrav, 1980) point out that subjectivist evaluation often leads to varying, sometimes contradictory, conclusions that defy reconciliation because that which led to the conclusions is largely obscured within the. nonrep— licable procedures of the evaluator. Similarly, as House puts it,

Critics of the phenomenologist epistemology note that there is often confusion over whose common sense perceptions are to be taken as the basis for understanding. Furthermore, if one takes everyday understanding as the foundation of inquiry, does one not merely reconstruct whatever ideologies, biases, and false beliefs already exist? How can one distinguish causal determinants and regularities, the strength of the positivist

epistemology, from perceived beliefs? How can one evaluate conflicting interpretations? Phenomenology provides no way of doing so. (House, 1980, p. 254)

And so the dialogue continues. The objectivists depend upon replicable facts as their touchstone of truth, whereas subjectivists depend upon accumulated experience as their way to understanding. Although both epistemologies carry within them "tests" that must be met if their application is to be viewed as trustworthy, they lead to very different evaluation designs and methods, giving rise to much of today's diversity in evaluation approaches.

Utilitarian Versus Intuitionist—Pluralist Evaluation

House (1976, 1983a) has also made a distinction closely related to that of objectivism and subjectivism, namely utilitarian versus intuitionist—plurahst evaluation. Although this is a distinction concerning principles for assigning values, not

epistemology, utilitarian and intuitionist—pluralist evaluation approaches parallel the objectivist and subjectivist episcemologies outlined above.

Utilitarian Evaluation. Utilitarian approaches determine value by assessing the *overall* impact of an educational program on those affected. These approaches have tended to follow objectivist epistemology. In his treatise on "justice in evaluation," House (1976) suggests that utilitarian evaluaticn accepts the value premise that the greatest good is that which will benefit the greatest number of individuals. Thus, "properly speaking, utilitarianism refers to the idea of maximizing happiness in society" (House, 1983a, p. *49)*. There is a single, explicitly defined ethical principle operative. As a result, the evaluator will focus on total group gains by using average test scores or some ,other common index of "good" to identify the "greatest good for the greatest number." The best educational programs are those that produce the greatest gains on the criterion or criteria selected to determine worth. Statewide assessment programs and large—scalecom— parative evaluations are utilitarian in nature. Most utilitarian—evaluation approaches lend themselves to use by governments or others who mandate and/or sponsor evaluation studies for which managers and public program administrators are the major audiences.

intuitionist-Pluralist Evaluation. At the opposite end of the continuum are intuitionist—pluralist approaches to evaluation, which are based **on the idea that** value depends upon the impact of the program on *each* individual. These approaches have tended to follow subjectivist epistemology. Here the value position is that the greatest good requires the attention to each individual's benefit. Thus,

The ethical principles are not single in number nor explicitly defined as in utilitarian ethics. There are several principles derived from intuition and experience, but no set rules for weighting them. This captures another meaning of ethical subjectivism—that the

ultimate criterion of what is good and right are individual feelings or apprehensions. (House, 1983a, p. 50)

This approach leads to a focus on the distribution of gains by individuals and subgroups (for example, ethnic groupings). There can be no common index of "good," but rather a plurality of criteria and judges, and the evaluator is no longer an impartial "averager" but a portrayer of different values, and needs. Data may be test scores, but intuitionist—pluralist evaluators often prefer data from personal interviews and testimonials of program participants. Weighing and balancing the many judgments and criteria inherent in this approach is largely intuitive, and there are no algorithms to help reduce complex evaluative information to any unequivocal recommendation. The perceived merit or worth of an educational program depends largely on the values and perspectives of whoever is judging, and each individual or-constituent group is a legitimate judge. "Likewise, the subjective utility of something is based on personal judgment and personal desires. Each person is the best judge of events for himself" (House, 19$3a, p. 56). Within limits of feasibility, most inrutionist—pluralist cvalutions try to involve as

"judges" all individuals and groups who arc affected by the program being evaluated, rather than leaving decisions and judgments to governmental sponsors and high—level administrators—as is typically the case with utilitarian evaluation.

The Impact of Philosophical Differences

Evaluators have tended to line up along the several continua described above, or worse, have become polarized in "either-or" dichotomies. Talmage sees this debate over epistemology as a major cause of rifts that permeate the field of evaluation. Speaking of such philosophical differences, Talmage said, "Program evaluation has been greatly affected by this schism: what is considered 'acceptable' evaluation research often depends upon the position taken regarding one or another of these continua" (Talmage, 1982, p. 596). Yet, although differences in philosophy have led to alternative views of evaluation, the philosophical differences are not incompatible. Multiple approaches to describing objects of study, drawn from both objectivist and subjectivist traditions, have been used in the same evaluations to achieve important goals (see, for example, Stake & Easely, 1978; Sanders & Sonnad, *1982).*

In choosing a philosophical orientation, evaluators need to consider (1) the credibility of results reported to evaluation clients, (2) the need for exploration when studying unknown phenomena, (3) the importance of understanding or explaining findings, (4) the need to be sensitive to emerging or hidden issues during the evaluation, and, of course, (5) the importance of thoroughly addressing questions posed by the client (that is, meeting the client's expectations) when planning an evaluation. We recognize the right of any evaluator to subscribe totally to the assumptions and premises of one particular ideology. Yet few evaluators who succeed in a wide range of evaluation settings can afford to consider philosophical ideologies as "either—or" decisions. The purist view that looks noble in print yields to practical pressures demanding that the evaluator use appropriate methods based on an epistemology that is rightfrr *that evaluation,* or even multiple methods based on alternative epistemologies within the same evaluation.20 It is important to know, however, the assumptions and limitations of methods that are drawn from different world views about evaluation.

METHODOLOGICAL BACKGROUNDS AND PREFERENCES

Different philosophical assumptions about knowledge and value give rise naturally to different evaluation methods. Indeed, evaluation philosophy and methodology are so closely intertwined that we might well have discussed both together in the previous section. But we believe it useful to examine separately two methodological issues that have influenced greatly the conduct of evaluation studies: (1) quantitative versus qualitative inquiry, and (2) the difficulty encountered by evaluators in working across disciplinary and methodological boundaries.

Quantitative and Qualitative Evaluation

Much has been said in recent years about quantitative and qualitative evaluation, as evaluators have struggled to sort out the relative utility of these two distinct approaches. To understand this distinction more fully, we must refer to the history of its evolution. Because so many people serving in evaluation roles during the late 1950s and 1960s were educational and psychological researchers, it is not surprising that the experimental tradition quickly became the most generally accepted evaluation approach. The work of Campbell and Stanley (1963, 1966) gave enormous impetus to the predominance of experimental or quasi-experimental approaches. Although some evaluators cautioned that correct use of the experimental model in largely uncontrollable classroom settings might not be feasible, the elegance and precision of the experimental method led most educational evaluators to view it as the ideal.

Not all program evaluators were enamored with the use of traditional quantitative methods for program evaluations, however, and their dissatisfaction led to a search for alternatives. Qualitative and naturalistic methods, largely shunned by most educational evaluators during the 1960s as unacceptably "soft," gained wider acceptance in the 1970s and thereafter as proposals for their application to program evaluations were made by Parlert and Hamilton (1976), Stake (1978, 1980), Eisner (1976, 1979b), Guba and Lincoln (1981), and others.

The rise in popularity of qualitative inquiry methods in education has been noted •by social and behavioral scientists concerned with the study of education (see Bogdan & Bikkn, 1982; Cook & Reichardt, 1979; Rist, 1980). Bogdan and Biklen speak on the spectacular increase in acceptability of qualitative techniques in educational research:

• A field once dominated by measurement, operationalized definitions, variables, and empirical fact has had to make room for a research approach gaining in popularity, one that emphasizes inductive analysis, description, and the study of people's perceptions

dependence on qualitative methods for studying various educational issues is growing qualitative research in education has, or will soon, come of age. (Bogdan **&** Biklen. 1982, p. xiii)

These same trends are no less true for educational evaluation; if anything, qualitative techniques have gained favor more quickly there.

Before proceeding further, we should delineate more clearly the differences between qualitative and quantitative methods. We find descriptions by Schofield and Anderson (1984) most useful for this purpose.21 In their view, *qualitative* inquiry generally (a) is conducted in natural settings, such as schools or neighborhoods; (b) utilizes the researcher as the chief "instrument" in both data—gathering and analysis,. . . (c) emphasizes "thick description," that is, obtaining "real," "rich," "deep." data which illuminate everyday patterns of action and meaning from the perspective of those being

studied. . „ (d) tends to focus on social processes rather than primarily or exclusively on outcomes, (e) employs multiple data—gathering methods, especially participant—

observation and interviews, and (fl uses an inductive approach to data analysis, extracting its concepts from the mass of particular detail which constitutes the data base.

By Contrast, these authors state that *quantitative* inquiry generally

focuses on the testing of specific hypotheses that are smaller parts of some larger theoretical perspective. This approach follows the traditional natural science model more closely than qualitative research, emphasizing experimental design and statistical methods of analysis. Quantitative research emphasizes standardization, precision, objectivity, and reliability of measurement as well as replicability and generalizability of findings. Thus, quantitative research is characterized not only by a focus on producing numbers but on generating numbers which are suitable for statistical tests. (Schofield & Anderson, 1984, pp. 8—9)

Although subtle differences exist, other terms used to describe very similar methodological distinctions should be noted here. Guba and Lincoln (1981) contrast the *naturalistic* and *scientific* paradigms. The dichotomy of *subjective* versus *objecrive* methodology is similar to chat of qualitative and quantitative methods as defined above. We should note that all of the dichotomies are largely artificial and should more accurately be thought of as different ends of a continuum (Goetz & LeCompte, 1981), although we will continue to use them dichotomously for simplicity, at least for now.

During the 1960s, sharp disagreements developed between proponents of the newer qualitative approaches and adherents to the more broadly accepted quanti.tative methods. And the 1970s were marked by debates as the two schools of thought struggled for ascendancy (see, for example, the Page and Stake, 1979, debate). Although some who favor qualitative methods are concerned that the sudden popularity and apparent simplicity of this approach have attracted innocents who employ qualitative inquiry without understanding of its complexity or the competence it demands of its user (Schofield & Anderson, 1984), most advocates are delighted by its increasing acceptance and are quick to attack weaknesses in quantitative inquiry. Those who favor quantitative methods are, for the most part, distressed by the shift toward qualitative inquiry (notwithstanding the fact that quantitative work is still the dominant approach to educational inquiry, as even casual reading of the most influential journals in education and related areas will reveal). Critics of qualitative evaluation often complain about the subjectivity of many qualitative methods and techniques, expressing concern that evaluation has abandoned objectivity in favor of inexpertly managed subjectivity.

In short, the last two decades have been marked by acrimony between these seemingly irreconcilable methodological persuasions.

Recently, however, the dialogue has begun to move beyond this debate, with analysts increasingly discussing the benefits of othniethods within an educational evaluation study (for example, Cook & Reichardt, 1979; Worthen, 1981; and the especially useful summary by Madey, 1982). Stone's (1984) comment about educational research seems to extend to evaluation as well:

Today in educational research,. . .the trend is methodological pluralism and eclecticism. Many formerly—devout quantitative researchers are now trying their hands at qualitative inquiry. The vigorous quantitative/qualitative debate, if not dead, is somehow buried. (Stone, 1984, p. 1)

Yet even here there is not agreement, and scholars hold very disparate views of how far the integration of the qualitative and quantitative paradigms can be carried. Some, like J. K. Smith (1983), see such fundamental differences that they believe there is little hope of meaningful integration. Others, like Guba and Lincoln (1981), acknowledge that complementarity might be *possible* on some dimensions, yet they opt for one paradigm and seem pessimistic about healing the schism:

Can one assume both singular reality and multiple realities at the same time? How can one believe in insularity between the investigator and the object of his investigation, while also allowing for their mutual interaction? How can one work simultaneously toward the development of nomothetic and idiographic science? (Guba & Lincoln, 1981, p. 77)

There is a growing chorus, however, of those who are optimistic about the fruitfulness of a complementary approach. For example, Howe (1985) has argued persuasively that adherence to rigid epistemological distinctions between qualitative and quantitative methods is, in itself, nothing more than a dogma held over from logical positivism.

Perhaps Schofield and Anderson (1984) state this position best when they say that

recent years have seen a number of important statements which argue against the traditional view that qualitative and quantitative work are based on fundamentally different paradigms and are thus competing and irreconcilable ways of approaching research.... Scholars of this persuasion, many of whom have been deeply involved with evaluation research in the field of education, argue that the distinction between qualitative and quantitative research is a matter of degree rather than of a basic difference which creates an unbridgeable chasm between the two camps..

Reichardt and Cook (1979) argue that method-type is not irrevocably linked to paradigm—type and that the qualitative and quantitative paradigms are neither as rigid nor as incompatible as is commonly assumed. For example, they argue that all research has important subjective elements and that the characterization of quantitative research as objective and of qualitative research as subjective overdraws the distinction between the approaches in numerous ways....

If qualitative and quantitative methods are not rooted in opposite and irreconcilable paradigms but rather are both more or less subjective, more or less valid and the like, there is no reason why they can not be utilized simultaneously. In fact, a number of scholars have recently argued not only that quantitative and qualitative approaches *can* be utilized jointly but that they *should* be so utilized.... The basic argument behind this position is that these two research strategies tend to have complementary strengths.... Qualitative research. . .is weak where experimental and other quantitative designs are often strong and strong where such designs are frequently weak. Specifically, qualitative research is not generally able to specify causal connections with the degree of certainty or precision that many quantitative strategies can. However, it is ideally suited to suggesting

ideas about soëial processes. **to exploring** the context in which the phenomena under investigation occur, and to capturing with both vividness and subtlety the perceptions of the itidividuals being studied. (Schofield & Anderson, 1984. pp. 12—13, 16—17)

We view quantitative and qualitative methods as compatible, complementary approaches in evaluation of educational programs. We have little interest in extending what we believe to be. the relatively meaningless arguments that favor quantitative methods over qualitative, or vice versa. We view both forms of inquiry as appropriate, depending on the purpose and questions for which the study is conducted. Our bias for an ecumenical resolution of this issue permeates and influences this book in ways we hope will prove useful. We echo Stone's conclusion, if only hopefully, that the quantitative/qualitative debate, if not dead, is somehow buried. But if not, we would suggest that the energy of educational scholars and practitioners still being expended in such debate be more productively channeled into conceptualizing and testing procedures for effective integration of quantitative and qualitative methodologies, an area in which there is still very little guidance.

Disciplinary **Boundaries and Evaluation Methodology**

It is ironic that in a field with such a rich array of alternative evaluation approaches. there still exists a tendency to fall prey to the "law of the instrument" fallacv,' rather than adapting or inventing evaluation methods to meet our needs. Our grasp of evaluation still seems partial and parochial, as may be expected in a young field. But it is unfortunate that we seem to carry with us into a new field the methodological allegiances we developed through earlier studies. Too often we fail to encourage methodological flexibility, unthinkingly adopting a single- minded perspective that can answer only questions stemming from that perspective. Today's typical evaluation studies depend largely on methodology adapted from agronomy, some fields in psychology, and to a limited extent, sociology.' Those who interpret this statement as critical of these fields have missed the point, for they are esteemed disciplines, with methodologies well suited to pursue research questions within their respective spheres of inquiry. Rather, the point is that evaluation is not a discipline but merely a social process or activity aimed at determining the yalue of certain materials, programs, or efforts. As such, it necessarily cuts across disciplines, and evaluators are thus denied the luxury of remaining within any single inquiry paradigm.

It has been argued previously (see Chapter 3) that evaluation, unlike research, cannot fix the boundaries of its own inquiry, that evaluation questions are set by clients' needs and might be framed so as to require the tools of several disciplines tb answer them. It was asserted that evaluators would need to have the flexibility to use econometrics to collect one type of data, psychometrics for another, socio— metrics for a third, and so forth. Yet evaluators often go about the business of evaluation using their preferred methods and drawing little if at all on the alternative paradigms that may be more relevant to the evaluation problems at hand, It

is not an easy thing to shuck *off* the conceptual shackles forged by experience. It is harder yet to expect that busy evaluators will interrupt an evaluation study to set *off* on an intellectual expedition into the terra incognita of another discipline to discover new methods and techniques perhaps more relevant to the problem at hand than they currently possess. And advising evaluators to be methodologically interdisciplinary sounds somewhat hollow in the absence of graduate programs designed specifically to assist evaluators-to-be in learning how they might do so.

Several writers have recognized the implications of evaluators' methodological preferences. For example, Talmage (1982) divides evaluators into four groups according to methodology: experimentalists, eclectics, describers, and benefit—cost analysts. Table 4.1 summarizes how, in her view, these four methodological positions relate to differences on several other dimensions; an excellent discussion of these differences appears in her original work.

Anderson and Ball (1978) also address how evaluators' predispositions and preferences on both philosophical and methodological dimensions led to differing designs, data collection and analysis methods, and interpretive techniques. Table 4.2 summarizes briefly their views, which are discussed in more detail in their original work.

More examples of alternative methodological preferences and their effect on evaluation studies could be given, but these should suffice. The increasing variety of methodological perspectives gaining legitimacy in educational evaluation is not only increasing the variety of ways evaluations are designed and conducted, but also is adding richness of perspective to a field still too young to opt for any single, ideal evaluation paradigm.

DIFFERENT METAPHORS OF EVALUATION

The importance of metaphors in evaluation has become increasingly clear during the past decade.24 Worthen (1978) described the rationale that led the Northwest Regional Educational Laboratory to launch a federally supported research effort to identify metaphors from other disciplines that might lead to useful new evaluation methodologies in the field of education:

If I may use a metaphor, we have proposed. . .a planned expedition into other fields to find and capture those methods and techniques that might have relevance for educational evaluation and. . . domesticate them so they will become tractable for our use. Again.

limited resources will allow us to explore only so far, so we need to identify early those areas which appear most likely to coniain good methodological candidates for domestication. (Worthen, 1978. p. 3)

Continued and enhanced greatly under the leadership of N. L. Smith (198th, 1981c), and with conceptual contributions by Guba (1978b), this research effort has examined the possibility of using a variety of metaphors, such as investigative journalism, photography, storytelling, philosophical analysis, and literary criticism, to mention only a few. Although several of these metaphors have proven of limited use for educational evaluation, others have yielded many useful new methods and techniques for evaluators in educational settings.

TABLE 4.1

Four Methodological Approaches in Program Evaluation

	Experimentalists	*Eclectics*	*Describers*	*Benefit-Cost Analyzers*
	Cook and Campbell (1979) Riecken and Boruch (1974) Rivlin and Timpane (1975)	Bryk (1978) Cronbach and others (1980) R. S. Weiss and Rein (1972)	Parlett and Hamilton (1977) Patton (1980) Stake (1975)	Haller (1975) Levin (1975) Thompson (1980)
Philosophical base	Positivist	Modified positivist to pragmatic	Phenomenological	Logical/Analytic
Disciplinary base	Psychology	Psychology; sociology; political science	Sociology; anthropology	Economics; accounting
Focus of methodology	Identify causal links	Augment search for causal links with process and contextual data	Describe program holistically and from perspective of the participants	Judge worth of program in terms of costs and benefits
Methodology	Experimental and quasi-experimental designs	Quasi-experimental designs; case studies; descriptions	Ethnography, case studies, participant observation triangulation	Benefit-cost analysis
Variables	Predetermined as input-output	Predetermined plus emerging	Emerging in course of evaluation	Predetermined
Control or comparison group	Yes	Where possible	Not necessary	Yes
Participants' role in carrying out evaluation	None	None to interactive	Varies (may react to field notes)	None
Evaluator's role	Independent of program	Cooperative	Interactive	Independent of program
Political pressures (internal–external)	Controlled in design; or ignored	Accommodated	Describe	Ignore
Focus of evaluation report	Render "go/no go" decision	Interpret and recommend for program improvement	Present holistic portrayal of program in process	Render judgment

Source: Talmage, 1982, p. 600. Harold E. Mitzel, Editor in Chief

TABLE 4.2

Predispositions and Preferences of Evaluators (Including Examples of Design, Measurement, Analysis, and Interpretation Preferences Associated with the Principal Dimensions)

	Phenomenological	*Behavioristic*
Design	Clinical or case study	Experimental or quasi-experimental design
Measurement	Subjective measurement methods, content analyses, self-reports	Objective measurement methods, tests, systematic observations
Analysis	Descriptive statistics and nonparametric techniques	Inferential statistics
Interpretation	Judgmental, value-laden	Nonjudgmental
	Absolutist	*Comparative*
Design	One-group design	Experimental or quasi-experimental design with comparison group(s)
Analysis	Within-group analysis	Between-group analysis
Interpretation	Standard-referenced	Comparison-group referenced
	Independent	*Dependent*
Measurement	Goal-free measures	Measures tailored to program goals
Interpretation	Nonclient-oriented	Goal-referenced, client-oriented
	Pragmatic	*Theoretical*
Design	Widely varying	Experimental or quasi-experimental design (hypothesis testing)
Measurement	Ad hoc measures, records	Established measures, construct validity emphasized
Analysis	Widely varying	Inferential statistics
Interpretation	Program-specific conclusions, little generalization (ideographic)	Hypothesis confirmation, generalization (nomothetic)
	Narrow Scope	*Broad Scope*
Measurement	Few and specific measures	Many and global measures
Analysis	Univariate contrasts	Multivariate analyses
Interpretation	Oriented toward component functioning	Oriented toward system functioning
	High Intensive	*Low Intensive*
Design	Repeated measurement occasions (longitudinal)	Infrequent measurement occasions (perhaps cross-sectional)
Measurement	Multitrait; multimethod (triangulation)	Survey tests

TABLE 4.2 CONTINUED

	High Intensive	*Low intensive*
Analysis	Multivariate analyses, including factor analyses	Univariate analyses, descriptive statistics
Interpretation	Generalization	Description
	Process	*Product*
Design	Repeated measurement occasions	Experimental or quasi-experimental design, infrequent measurement occasions
Measurement	Observations, logs, interviews	Tests
Analysis	Descriptive statistics	Inferential statistics
Interpretation	Recommendations for program improvement	Recommendations for program continuation, expansion, "accreditation"

Source: Anderson and Ball, 1978, pp. 122–123

One need not consciously seek metaphors in the way the Northwest Regional Educational Laboratory study did. Metaphors underlie and influence much of our thinking. Indeed, one reason for differing evaluation approaches is the different evaluation metaphors held by writers and practitioners. House (1983b) has demonstrated that much of our everyday thinking is metaphorical in nature and extends that point to argue that evaluation thought is also largely metaphorical. Further, he suggests that conflicts between existing evaluation schemes stem from differences in the underlying metaphors held by proponents of those schemes. For example, metaphoric conceptions of social programs equate those programs with industrial production (leading to metaphors based on machines, assembly lines, or pipelines) or with sports contests or games (leading to metaphors of targets and goals).

The influence of such metaphors on evaluation is obvious. For example, one who perceives evaluation as retrospective backtracking of a program to discover the causes of its outcomes is likely to use an approach that resembles forensic pathology, whereas one who holds a connoisseurial metaphor of evaluation will use an approach more akin to literary criticism. Yes, different metaphors account for much of the variation in evaluation approaches.

RESPONDING TO DIFFERENT NEEDS IN EDUCATION

In proposing new evaluation approaches, evaluation theorists have not only been influenced by their different methodological and metaphorical preferences or their different ways of looking at knowledge and how it is achieved. They have also been responding to different needs that they perceived, needs such as school administrators wanting better information for decision making, educators wanting a more systematic way to review school curricula, school personnel struggling for

of that which is evaluated (for example, knowledge of mathematics in evaluating a mathematics education program)—some evaluators (for example, Eisner, 1975, 1979a) see such expertise as the *sine qua non* of evaluation. Indeed, without such expertise, their evaluation approach would be futile. Other evaluators (for a better way to articulate the purposes of schooling, federal and state legislators monitoring resource allocation, and local stakeholders hoping to define issues in rational rather than emotional terms.

Various approaches were developed to address each need. In the aggregate, these different approaches help us comprehend the wide range of needs for evaluation in education. We must learn to identify what is useful in each approach when faced with a specific evaluation need, to use it wisely, and not to be distracted by irrelevant evaluation approaches constructed to deal with a different need.

PRACTICAL CONSIDERATIONS

We have traced how epistemological issues, methodological preferences, metaphoric views of evaluation, and different needs all contribute to the diversity of alternative evaluation approaches. Several practical issues also contribute to this diversity.

First, evaluators disagree about whether the intent of evaluation is to render a value judgment. Some (Edwards, Guttentag. & Snapper, 1975; Patton, 1978; and Weiss, 1977) are concerned only with the usefulness of the evaluation to the decision—maker and believe that he, not the evaluator, should render the value judgment. Others (for example, Scriven, 1973; Wolf, 1979; and Worthen & Sanders, 1973) believe the evaluator's report to the decision—maker is complete only if it contains a value judgment. Such differences in views have obvious practical implications.

Second, evaluators differ in their general view of the political roles of evaluation. MacDonald (1976) provides a political classification (described in greater detail in Chapter 10) to illustrate how political orientation affects the style of evaluation conducted: bureaucratic, autocratic, or democratic. Although MacDonald prefers the democratic, he provides a useful analysis of how these different political— orientation issues result in very different approaches.

Third, evaluators are influenced by their prior experience. Although most recent conceptual work in educational evaluation has come from academics in higher education, some of it emanates from practitioners in local and state education agencies. Many of the authors who have shaped our thinking about evaluation have been educational psychologists, specialists in educational and psychological tests and measurements, guidance counselors, school administrators, curriculum specialists, or philosophers. Each theorist has drawn from certain strengths, from experience with certain types of problems and processes in education, and from a way of looking at things that grew out of his or her professional education and career. In the aggregate, the different approaches represent nearly every way

imaginable of looking at education; individually, however, they represent limited perspectives.

Fourth, evaluators differ in their views about who should conduct the evaluation and the nature of the expertise that the evaluator must possess. Although this topic is too complex to be treated adequately in this chapter, an illustration might help. Considering one dimension of expertise—substantive knowledge about the content example, Worthen & Sanders, 1984) not only question the need for the evaluator to possess such expertise but also suggest that there may sometimes be advantages in selecting evaluators who are not specialists in the content of that which they ea]u Such differences in perspective lead to different approaches to evaluation.

Finally, evaluators differ even in their perception of whether it is desirable to have a wide variety of approaches to educational evaluation. Antonoplos (1977) and Gephart (1977) have lamented the proliferation of evaluation models and urged that an effort be made to synthesize existing modcls. Palumbo and Nachmias (1984) are less sanguine about the possibility of developing an ideal evaluation paradigm, but they propose that making the effort is worthwhile. Conversely, Raizen and Rossi (1981) have argued that the goal of attaining uniformity in evaluation methods and measures, proposed by some as a way to increase quality of information, cannot be attained at the present time without prematurely inhibiting needed development in the field of evaluation. Worthen (1977b) had earlier made a similar point in arguing that efforts to synthesize existing evaluation models would be dysfunctional, an argument that will be expanded later in Chapter 11. Regardless of which view you subscribe to, it is clear that either the inability to generate an idealistic evaluation model (after all, none has been forthcoming since the call for synthesis nearly a decade ago) or resistance to trading the diversity of models for a unified view accounts, at least in part, for the continued variety of approaches that confronts the evaluation practitioner.

THEMES AMONG THE VARIATIONS

Despite the diversity in evaluation approaches, commonalities do exist. Many individuals have attempted to bring order out of the chaos reflected in evaluation literature by developing classification schemes, or taxonomies. Each such effort selected one or more dimensions deemed useful in classifying evaluation ap-. proaches. But because evaluation is multifaceted and because it can be conducted at different phases of a program's development, the same evaluation model can be classified in diverse ways. depending on emphasis. Consider the following examples. Those who have published classification schema include Worthen and Sanders (1973), Popham (1975), Ross and Cronbach (1976), Stake (1975b), Curriculum Development Centre (1977), Stufflebeam and Webster (1980), Guba and Lincoln (1981), House (1983a), Madaus, Scriven, and Stufflebeam (1983), and Worthen (1984). All have influenced our thinking about the categorization of evaluation

approaches, but we have drawn especially on our own work and that of House in developing the schema proposed below.

A CLASSIFICATION SCHEMA FOR EVALUATION APPROACHES

We have chosen to classify many different approaches to evaluation into the six categories described below.

1. *Objectives-oriented* ***approaches,*** where the focus is on specifying goals and objectives and determining the extent to which they have been attained.

2. *Management-oriented approaches,* where the central concern is on identifying and meeting the informational needs of managerial decision—makers.

3. ***Consumer-oriented*** *approaches,* where the central issue is developing evaluative information on educational "products," broadly defined, for use by educational consumers in choosing among competing curricula, instructional products, and the like.

4. *Expertise-oriented approaches,* which depend primarily on the direct application of professional expertise to judge the quality of educational endeavors.

5. *Adversary-oriented approaches,* where planned opposition in points of view of different evaluators **(pro** and con) is the central focus of the evaluation.

6. *Naturalistic and participant-oriented approaches.* where naturalistic inquiry and involvement of participants (stakeholders in that which is evaluated) are central in determining the values, criteria, needs, and data for the evaluation.

These six categories seem to us to distribute (though not equally) along House's (1983a) dimension of utilitarian to intuitionist—pluralist evaluation, as shown in Figure 4.1 below.

Placement of individual evaluation approaches within these six categories is to some degree arbitrary. Several approaches are multifaceted and include characteristics that would allow them to be placed in more than one category; for convenience we have decided to place such approaches in one category and only reference in other chapters, where appropriate, their other features. Our classification is based on what we see as the driving force behind doing the evalj.iation—the major questions to be addressed and/or the major organizer(s) that underlie each ·approach (for example, objectives, or management decisions). Within each category, the approaches vary by level of formality and structure, some being relatively well developed philosophically and procedurally, others less developed, it should be noted that these frameworks deal with conceptual approaches to evaluation, not techniques; discussion of the many techniques that might be used in educational evaluations is reserved for Part Three of this book.

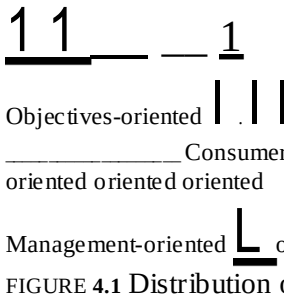


FIGURE **4.1** Distribution of Six Evaluation Approaches on the Dimension of Utilitarian to ntuitionist—PluraIist Evluation

APPLICATION EXERCISE

1. Think about how you would approach evaluation. Describe the steps you think you would follow. Then, analyze your approach according to your philosophical and-methodological *preferences*. Explain how your background and what you would be evaluating could have affected your approach. Describe other things that might have affected your approach to evaluation.
2. Identify an educational program you would like to see evaluated. List some qualitative evaluation methods that could be used. Now list some quantitative methods that you see as appropriate. Discuss whether it would be appropriate to combine both methods within the same study, including reasons for your conclusion.

SUGGESTED READINGS

GUBA, E. G., & LINCOLN, Y. S. (1981). *Effeait'e evaluation.* San Francisco: Jossey-Bass.

Housa, *E*. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage.

HousE, E. R. (1983a). Assumptions underlying evaluation models. In G. F. **MADAUS,** M. SCRIVEN, & D. L. **STUFFLEBEAM** (Eds.), *Evaluation models.* Boston: Kluwer-Nijhoff.

Housr, E. R. (Ed.). (1983b). *Philosophy of evaluation*. New Directions for Program Evaluation, No. 19. San Francisco: Jossey-Bass.

The Tylerian Evaluation Approach

Tyler's approach to evaluation was developed and used during the Eight Year Study of the late l930s (Smith & Tyler, 1942). Travers (1983) did note, however, that an earlier work, Waples and Tyler's *Research Methods and Teacher Problems* (1930), set the stage for Tyler's later achievements in evaluation.

Tyler conceived of evaluation as the process of determining the extent to which the educational objectives of a school program or curriculum are actually being attained. His approach to evaluation followed these steps:

1. Establish broad goals or objectives.
2. Classify the goals or objectives.
3. Define objectives in behavioral terms.
4. Find situations in which achievement of objectives can be shown.
5. Develop or select measurement techniques.
6. Collect performance data.
7. Compare performance data with behaviorally stated objectives.

Discrepancies between performance and objectives would lead to modifications intended to correct the deficiency, and the evaluation cycle would be repeated.

Tyler's rationale was logical, scientifically acceptable, readily adoptable by educational evaluators (most of whose methodological upbringing was very compatible with the pretest—posttest measurement of student behaviors stressed by Tyler), and had great influence on subsequent evaluation theorists.

Goodlad (1979) pointed out that Tyler advocated the use of general goals to establish purposes, rather than prematurely becoming preoccupied with formulating behavioral objectives. Of course, the broad goals for any activity eventually require operational definitions so that appropriate measurement devices and settings can be selected. Tyler's belief was that educators primarily needed to discuss the importance and meaning of general goals of education. Otherwise, in Goodlad's words, the premature specification of behavioral objectives would result in objectives that, "could only be arbitrary, restrictive, and ultimately dysfunctional" (Goodlad, 1979, p. 43).

Tyler described six categories of purpose for American schools (Goodlad, 1979). They were (1) acquisition of information; (2) development of work habits and study skills; (3) development of effective ways of thinking; (4) internalization of social attitudes, interests, appreciations, and sensitivities; (5) maintenance of physical health; and (6) development of a philosophy of life.

Over the years, educators have refined and reformulated the purposes of schooling into various forms. Two recent publications provided statements that reflect the thinking of the past 50 years.

First, the Evaluation Center at Western Michigan University developed a *Handbook of Educational Variables* in cooperation with the Toledo, Ohio, Public Schools (Nowakowski and others, 1985). The *Handbook* divided elementary and secondary student development into these seven categories:

1. Intellectual
2. Emotional

3. Physical and Recreational
4. Aesthetic and Cultural
5. Moral
6. Vocational
7. Social

Each one of these categories was analyzed in detail too extensive to reproduce here. Such a resource exemplifies the extent to which Tyler's approach to evaluation has been refined.

A second resource that specifies the purposes of schooling in the United States is Goodlad's (1979) list of 12 goal areas for schools, which includes the following major categories:

1. Mastery of basic skills or fundamental processes
2. Career education/vocational education
3. Intellectual development
4. Enculturation
5. Interpersonal relations
6. Autonomy
7. Citizenship
8. Creativity and aesthetic perception
9. Self-concept
10. Emotional and physical well-being
11. Moral and ethical character
12. Self—realization

Goodlad stressed that evaluation and improvement of American schools cannot make much headway until these purposes have been discussed, accepted. operationally defined, and monitored, It should be clear that a single standardized test of achievement of basic skills provides insufficient data to evaluate our schools. Yet the use of standardized test results is still the most common form of school evaluation discussed in the popular media today.

Tyler stressed the importance of screening broad goals before accepting them as the basis for evaluating an activity. The screen through which potential goals should be filtered includes value questions derived from three sources: philosophical (the nature of knowledge); social (the nature of society); and pedagogical (the nature of the learner and the learning process). Scriven (1967) reiterated the need to evaluate the purposes of any activity as a part of evaluating the activity and its consequences.

The question of how specifically to evaluate goals and objectives was addressed by Sanders and Cunningham (1973, 1974). Their approach was to consider both logical and empirical methods for goal evaluation. ***Logical*** methods included:

1. Examining the cogency of the argument or rationale behind each objective. If there are no justifiable reasons for a goal or objective, it cannot have much value. The ***need*** for accomplishing the goal or objective is a critical consideration.
2. Examining the consequences of accomplishing the goal or objective. By projecting logically the consequences of achicvmg a goal, both strengths and

weaknesses in compcting goak may he revealed. Criteria such as utility and feasibility (cost, acceptability, political palatability, training, or other require-. ments) of the goal or objective could be used here. A search of educational literature may reveal the results of past attempts to achieve certain goals or objectives.

3. Considering whether higher—order values such as laws, policies, fit with existing practices, moral principles, the ideals of a free society, or the Constitution, to see if a goal or purpose is required by or will conflict with such values.

Empirical methods for evaluating goals or objectives included:

1. Collecting grou data to describe judgments about the value of a goal or objective. Surveys arc the most common form of gathering information about a group's value position.
2. Arranging for experts, hearings, or panels to review and evaluate potential goals or objectives. Specialists can draw from knowledge or experience that may not be otherwise available. Their informed judgment may be very different from the group value data that surveys would produce.
3. Conducting content studies of archival records, such as speeches, minutes, editorials, or newsletters. Such content analyses may reveal value positions that conflict with, or are in support of, a particular goal or objective.
4. Conducting a pilot study to see if the goal is attainable and in what form it may be attained. If no prior experience is available when evaluating a purpose or goal, it may be advisable to suspend judgment until some experience has been gained. Once a broad goal has been made operational, or activities directed toward attaining the goal have been tried, it may take on a different meaning from that which it had in earlier discussions.

Several evaluation approaches developed in education during the late 1960s and early 1970s used goals or objectives as a central .focus in the evaluation procedure. These approaches may be seen, therefore, as further refinements of Tyler's approach. Most noteworthy of these objectives—referenced evaluation approaches were those developed by Metfessel and Michael (1967), Hammond (1973) and Provus (1969, 1971; Yavorsky, 1976). They are noteworthy because they added new insights into how educational programs may be studied within the Tylerian tradition.

Metfessel and Michael's Evaluation Paradigm

An early approach to evaluation suggested by Metfessel and Michael (1967) was heavily influenced by the Tylerian tradition. Eight steps in the evaluation process were proposed as follows:

1. Involve the total school community as facilitators of program evaluation.
2. Formulate cohesive model of goals and specific objectives.
3. Translate specific objectives into a communicable form applicable to facilitating learning.in the school environment.

4. Select or construct instruments to furnish measures allowing inferences about program effectiveness.
5. Carry out periodic observations using content—valid tests, scales, and other behavioral measures.
6. Analyze data using appropriate statistical methods.
7. Interpret the data using standards of desired levels of performance over all measures -
8. Develop recommendations for the further implementation, modification, and revision of broad goals and specific objectives.

One of the primary contributions of Metfessel and Michael was in expanding the educational evaluator's vision of alternative instruments that might be used to collect evaluation data. Interested readers will find their lists of alternative instru.ments for data collection (Metfessel & Michael, 1967; Worthen & Sanders, 1973, pp. 276—279) to be a valuable guide.

Hammond's Evaluation Approach

Hammond was interested not only in determining whether goals or objectives were achieved but also in finding out why some educational innovations failed while others succeeded. To help the evaluator search for factors that influence the success or failure of any educational activity, Hammond developed a three- dimensional cube (Hammond, 1973) for use in describing educational programs and organizing evaluation variables (see Fig. 5.1). Hammond called his cube a "structure for evaluation."

The three dimensions of the cube are:

1. ***Instruction:*** characteristics of the educational activity that is being evaluated.
a. ***Organization:*** Time, scheduling, course sequences, and organization of the school, iiThluding vertical (graded or ungraded) and horizontal (self— contained, cooperative teaching, or departmentalized) organization.
b. ***Content:*** Topics to be covered.
c. ***Method:*** Teaching activities, types of interaction (for example, teacher— student, media—student), teaching/learning theory.
d. ***Facilities:'*** Space, equipment, expendable materials.
e. ***Cost:*** Funds required for facilities, maintenance, personnel.
2. ***Institution:*** Characteristics of individuals or groups involved with the educational activity being evaluated.
a. ***Student*** (column I of the cube): Age, grade level, sex, family background, social class, health, mental health, achievement, ability, interests.
b. *Teacher, administrator, edutational specialist* (columns 2, 3, and 4 of the cube): For each role, one might attend to age, sex, race or religion, health, personality, educational background and work experience, pertinent personal characteristics (work habits).
c. *Family* (column 5 of the cube): Degree of involvement with the activity being evaluated, general characteristics such as culture or language, family

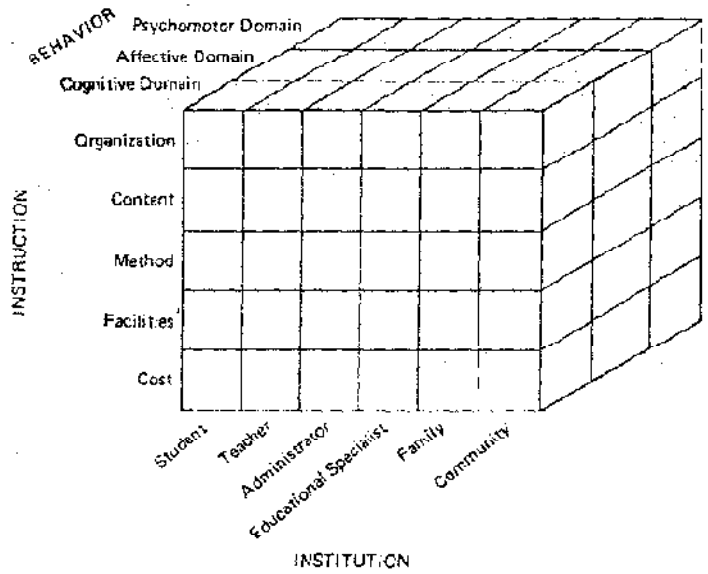


FIGURE 5.1 Structure for Evaluation

Source: Hammond, 1973, p. 158.

size, marital status, income, educational level, affiliations (for example, religion, politics, union).

d. *Community* (column 6 of the cube): Geographical setting, history, demographics, economic characteristics, social and political characteristics.

3. *Behavioral Objectives*: Objectives of the educational activity being evaluated.
   a. *Cognitive objectives*: Knowledge and intellectual skills.
   b. *Affective objectives*: Interests, attitudes, feelings, and emotions.
   c. *Psychomotor objectives*: Physical skills, coordination.

Hammond's cube is made up of 90 potentially useful cells. Any cell may be examined to determine the types of evaluative questions that might be generated. For example, an evaluator might examine the cell formed by the interaction of *content* (from the Instruction dimension), *teacher* (from the Institution dimension), and *affective objectives* (from the Behavioral Objectives dimension). What questions does this configuration of factors suggest? Here are just a few of many possible examples:

- How well are teachers using the subject matter of this program in achieving its (or their) affective objectives?
- How do teachers feel about the content and the program's affective objectives?
- Is the content of the program sufficient for teachers to accomplish their affective objectives?
- Are teachers prepared to use this subject matter to accomplish the program's affective objectives?

Hammond's evaluation structure is a valuable heuristic tool the evaluator can use in analyzing the successes and failures of an educational activity in achieving its objectives. If all cells in the cube were equally pertinent to a given evaluation study, the evaluator would have 90 cells to use for generating evaluation questions! Generating and responding to so many questions would be a monumental but extremely helpful task. Often, however, many of the cells prove irrelevant to a given evaluation, and frequently only a few cells apply.

Beyond the introduction of his cube, Hammond's approach to evaluation deviated little from the Tylerian approach, proposing the following steps in evaluating school programs:

1. Defining the program.
2. Defining the descriptive variables (using hi "cube").
3. Stating objectives.
4. Assessing performance.
5. Analyzing results.
6. Comparing results with objectives.

Provus's Discrepancy Evaluation Model

Another approach to evaluation in the Tvlerian tradition was developed by *Provus,* who based his approach on his evaluation assignments in the *Pittsburgh.* Pennsylvania, public schools. Provus viewed evaluation as a continuous infor— macion management process designed to serve as "the watchdog of program management" and the "handmaiden of administration in the management of program development through sound decision making" (Provus, 1973. p. 186). Although his was in some ways a management-oriented evaluation approach, the key characteristic of Provus's proposals stemmed from the Tylerian tradition. Provus viewed evaluation as a process of (1) agreeing upon standards (another term *used* in place of "objectives"2t'), *(2)* determining whether a discrepancy exists between the performance of some aspect of a program and the standards set for performance, and (3) using information about discrepancies to decide whether to improve, maintain, or terminate the program or some aspect of it. Provus called his approach, not surprisingly, the Discrepancy Evaluation Model.27

As a progiam or educational activity is being developed, Provus conceived of it as going through four developmental stages, to which he added a fifth optional stage:

1. Definition
2. Installation
3.. Process (interim products)
4. Product
5. Cost-benefit analysis (optional)

During the *definition,* or design, stage, the focus of work is on defining goals, processes, or activities, and delineating necessary resources and participants to carry out the activities and accomplish the goals. Provus considered educational programs to be dynamic systems involving inputs (antecedents), processes, and

outputs (outcomes). Standards or expectations were established for each. These standards were the objectives on which all further evaluation work depends. The evaluator's job at the design stage is to see that a complete set of specifications is, produced and that they meet certain criteria: theoretical and structural soundness.

At the *installation* stage, the program design or definition is used as the standard against which to judge program operation. The evaluator performs a series of congruency tests to identify any discrepancies between expected and actual installation of the program or activity. The intent is to make certain that the program has been installed as it had been designed. This is important because studies have found that teachers vary as much in implementing a single program as they do in implementing several different ones. The degree to which program specifications are followed is best determined through firsthand observation. If discrepancies are found at this stage, Provus proposed either changing the program definition, making adjustments in the installation (such as preparing a special in—service workshop), or terminating the activity if it appears that further development would be futile.

During the *process* stage, evaluation is to focus on gathering data on the progress of participants (for example, students) to determine whether their behaviors changed as expected. Provus used the term *enablin,g objective* to refer to those gains that participants should be making if program goals are to be reached. If certain enabling objectives are not achieved, the activities leading to those objectives are revised or redefined. The validity of the evaluation data would also be questioned. If the evaluator finds that enabling objectives are not being achieved, another option is to terminate the program if it appears that the discrepancy cannot be eliminated.

At the *product* stage, evaluation is to determine whether the *terminal objectives* for the program have been .achieved. Provus distinguished between immediate **Out-• comes,** or *terminal objectives,* and long—term outcomes, or *ultimate objectives.* He encouraged the evaluator to go beyond the traditional emphasis on end-of-program performance and make follow-up studies a part of evaluation.

Provus also suggested an optional fifth stage that called for *cost_benefit analysis* and comparison of results with similar cost analyses of comparable educational programs. In recent times, with funds for education becoming scarcer, cost—benefit analyses have become an essential part of almost all educational evaluations.

Thea Discrepancy Evaluation Model was designed to facilitate development of educational programs in a large public school system, and later it **was** applied to statewide evaluations by a federal bureau. A complex approach that works best in larger systems with adequate staff resources, its central focus is on use of discrepancies to help managers determine the extent to which program development is proceeding toward attainment of stated objectives. It attempts to assure effective program development by preventing the activity from proceeding to the next stage until all identified discrepancies have been removed. Whenever a discrepancy is found, Provus suggested a cooperative problem-solving process for program staff and evaluators. The process involved asking (1) Why is there a discrepancy? (2) What corrective actions are possible? and(3) Which corrective action is best? This

process usually required that additional information be gathered and criteria developed to allow rational, justifiable decisions about corrective actions (or ter.minations). This particular problem—solving activity was a new addition to the traditional objectives—oriented evaluation approach.

The evaluation approaches outlined here have been referred to not only as objectives—oriented evaluation approaches, the term we prefer, but also as "objectives—referenced" evaluations, "objectives-performance congruence" approaches, "performance congruence" models, and other similar terms. In each, assessment of the extent to which objectives have been attained is the central feature.

HOW THE OBJECTIVES-ORIENTED EVALUATION APPROACH HAS BEEN USED

The objectives-oriented approach has dominated the thinking and development of educational evaluation since the 1930s, both here in the United States and elsewhere, Its straightforward procedure of letting objectives achievement determine success or failure and justify improvements, maintenance, or termination of educational activities has proved an attractive prototype.

In the 1950s and early 1960s, curriculum evaluation and curriculum development procedures were based almost entirely on Tyler's conception of evaluation developed during the Eight Year Study. Taba, who worked with Tyler in the Eight Year Study, was one who influenced the field of curriculum development with an objectives-oriented model that included the following steps:

1. Diagnosis of needs
2. Formulation of objectives
3. Selection of content
4. Organization of content
5. Selection of learning experiences
6. Organization of learning experiences
7. Determining the "what" and "how" of evaluation (Taba, 1962, P. 12)

The technology of objectives—oriented evaluation was refined by Mager (1962), who went beyond simple insistence that objectives be prespecified in behavioral terms to insist that objectives must also contain the attainment levels desired by educators and the criteria for judging such attainment. Insistence on the use of behavioral objectives sparked a profession—wide debate that began in the l960s and still continues. Several educators (Gideonse, 1969; Popham, 1973a, 1973b, 1975) have championed the use of behavioral objectives, whereas others (for example, Atkin, 1968) contend that specification of behavioral objectives is not really helpful to curriculum development or sound curriculum evaluation.

Whether or not one believes behavioral objectives are useful, one cannot help being distressed by the mindlessness that ran rampant through education wherein educators were encouraged to state every intent—however trivial—in behavioral terms. In some schools, the staffs were spending so much time and energy stating everything they wanted to teach in behavioral terms that they hardly had time to teach. Training every teacher to use a recipe for translating every aspiration into a

behavioral objective wastes time and resources and distorts what education should be. This is especially true when teachers are used to writing objectives intended more for evaluation than for instruction, It is, after all, the evaluator who is

• supposedly skilled in the language of operationalization. We think the evaluator should take the following stance in working with program personnel: "Give me an objective in any form, just so I understand what your intent is. As an evaluator, I will translate your objective into behavioral terms and have you review my statement to make certain I have not distorted your intent." That approach makes more sense than trying to train all educators to be evaluators.

The pendulum obviously needed to move from the irresponsible position that educators do not need objectives because,, after all, they "know in their hearts they are right." But education moved too far in the opposite extreme when it spawned the religion of behaviorism and the disciples who applied it unintelligently. One can hardly oppose using instructional objectives and assessing their attainment, but the use of dozens or even hundreds of objectives for each area of endeavor, not uncommon a few years ago, amounted to a monopolization of educators' time and skills for a relatively small payoff. Although vestiges of this philosophy are still evident in some schools, the press for behavioral reductionism seems to have diminished. Had the push for behavioral objectives not been contained, disenchanted educators may well have refused to have anything to do with evaluation—an outcome that would have had serious consequences for education.

Though the debate has shifted from a focus on proper statement of objectives to that of how the objectives are to be measured, it still divides the field of evaluation. Bloom and Krathwohl were influential in refining the objectives-oriented approach to evaluation with their work on the previously discussed taxonomiés of educational objectives in both the cognitive (Bloom and others, 1956) and affective domains (Krathwohl and others, 1964). With the development of these taxonomies of objectives, curriculum specialists had powerful tools to aid them in using Tyler's approach. Bloom, Hastings, and Madaus (1971) also prepared a handbook for educators to use in identifying appropriate objectives for instruction in the subject matter taught in school and for developing and using measurement instruments to determine students' levels of performance in each subject. Cronbach (1963), who also worked with Tyler on the Eight Year Study, developed an approach to using objectives and associated measurement techniques for purposes of course and curriculum improvement.

But the blockbuster, in terms of expenditure, has been the objectives—referenced or criterion—referenced testing programs originated in the 1960s and 1970s by federal and state governments. The National Assessment of Educational Progress (NAEP) was originated in the mid-1960s under the leadership of Tyler. This federal program was designed to collect performance data periodically on samples of students and young adults in the essential subjects of American education. Great care was taken to select objectives generally accepted in this country as desirable achievements at the different stages of development measured (ages 9, 13, 17, and adult). Public reports have, since the mid—1960s, described the ability of Americans in these age groups to answer questions in subjects considered important.

Like those of the Eight Year Study, the instruments and objectives of NAEP have been made available to educators, but they have received limited use. Virtually every state has developed its own form of annual statewide testing, and many have generally followed the NAEP approach.

Derivative "objectives-oriented" movements in education, in the form of school accountability (Lessinger, 1970; Lessinger & Tyler, 1971). competency or minimum competency testing (Bunda & Sanders, 1979; Jaeger & Tittle, 1980; Madaus, 1983), objectives— and criterion-referenced collections and exchanges (Instructional Objectives Exchange, 1969; Clearinghouse for Applied Performance Testing, 1974), and federal project monitoring (such as the TIERS System for Title I projects developed by Tailmadge & Wood, 1976) appeared in the late 1960s and continue to be influential. For example, most states in the United States now have a competency testing program used to determine whether children have mastered the minimal objectives established for their respective grade levels. The tradition begun by Tyler over 50 years ago has had remarkable staying power.

STRENGTHS AND LIMITATIONS OF THE OBJECTIVES-ORIENTED EVALUATION APPROACH

Probably the greatest strength and appeal of the objectives-oriented approach to evaluation in education lies in its simplicity. It is easily understood, easy to follow and implement, and produces information that educators generally agree is relevant to their mission. This approach has stimulated so much technological development over the years that the processes of specifying objectives and developing or finding appropriate measurement procedures and instruments have been finely honed. The literature on objectives-oriented evaluation is extensive, filled with creative ideas for applying the approach in classrooms, schools, school districts, and beyond (Cronbach, 1963; Lindvall, 1964; Popham, Eisner, Sullivan, & Tyler, 1969; Metfessel & Michael, 1967; Bloom, Hastings, & Madaus, 1971: Morris & Fit zgibbon, 1978).

The objectives—oriented evaluation approach has caused educators to reflect about their intentions and to clarify formerly ambiguous generalities about educational outcomes (Mager, 1962). Discussions of appropriate educational objectives with the community being served have given objectives-oriented evaluation the appeal of face validity—the program is, after all, merely being held accountable for what its designers said **it** was going to accomplish, and that is obviously legici— mate.

As a result of the attention placed on this approach, tests have improved, and technically sound measurement practices have broadened to include unobtrusive (Webb, Campbell, Schwartz, & Sechrest, 1966) and non-paper-and-pencil evidence (Sanders & Sachse, 1977). These and other advances in the measurement of *ottcomes* in education may be tied to the outcome orientation of Tyler. These advances, added to the many instruments, objectives pools, and step-by—step guides that have been placed in the hands of educators by various projects, have

greatly expanded the available resources for educational evaluation during the twentieth century.

Useful as this approach to evaluation seems to its many adherents, critics have asserted that it (I) lacks a real evaluative component (facilitating measurement and assessment of objectives rather than resulting in explicit judgments of merit or worth), (2) lacks standards to judge the importance of observed discrepancies between objectives and performance levels, (3) neglects the value of the objectives themselves, (4) ignores important alternatives that should be considered in planning an educational program, (5) neglects transactions that occur within the program or activity being evaluated, (6) neglects the Context in which the evaluation takes place, (7) ignores important outcomes other than those covered by the objectives (the unintended outcomes of the activity). (8) omits evidence of program value not reflected in its own objectives, and (9) promotes a linear, inflexible approach to evaluation. Collectively, these criticisms suggest that objectives— oriented evaluation can result in tunnel vision that tends to limit evaluation's effectiveness and potential.

To some extent, the rather elaborate technology developed to support this evaluation approach makes its use appear seductively simple to novice evaluators only partially familiar with its philosophical and practical difficulties. The assumption that education is a technology—a body of techniques leading to prespecified means—has been criticized by Waks (1975), who pointed to potential problems with the philosophical underpinnings of this approach.

A recently passed law in one of our states is a classic example of poor use of the objectives—oriented evaluation approach. The General Assembly mandated that each school district should report once a year both to its local constituency and to the state the extent to which the district had achieved its stated goals and objectives. All a district had to do was to announce some general goals and specific objectives, carry out its program for a year, and at the end of the time report how well it had done on those goals and objectives. Although it is often important to know whether a district is attaining its stated objectives, such is not always the case. It depends largely on whether the goals were worth attaining in the first place. Some goals that are attainable are hardly worth the effort. Some goals are attained because they were set too low or had already been attained, not because the program was effective. The situation is almost analogous to that in which one needs to identify which children in a classroom are in good health and which are suffering from malnutrition, and height is considered a relevant indicator. There would be at least a measure of foolishness in asking each child to make his own tape measure, use it in measuring his height, and then report how well he has attained the height he desired to reach at that point or whether he is too tall or too short for his age.

A related difficulty lies in the frequent challenge of trying to ascertain the goals or objectives of many educational endeavors. Evaluators have found that the objectives listed on paper do not always match those in the minds of program staff As a result, their activities may conflict with or deviate from publicly stated

objectives, sometimes for good reasons. Professional educators will not become slaves to stated objectives if they believe alternative courses of action or goals are desirable. Such individuals tend to argue against strident and unthinking application of an objectives-oriented approach.

A related and perhaps more pervasive problem is the fact that many educators have not articulated objectives for their curricula or programs in any interpretable form. This is not to say they have no idea of where they are going or what they want to accomplish (although that is sometimes unfortunately the case), but rather that they are unaccustomed to thinking or speaking in "behavioral" language familiar to objectives—oriented evaluators. The evaluator may find it necessary to elicit clear statements of intent in an almost Rogerian fashion rather than translate those statements into behavioral terms as deemed necessary (not forgetting to ask those whose intentions they reflect if there have been distortions in the translation). Given the fact that many evaluators, lamentably, are not equipped by disposition or training to assist educators in this way, the objectives—oriented approach to evaluation frequently results in the verdict that a program cannot be evaluated, when the problem lies more with narrow understanding of the approach and/or the insistence that all educators must become experts in behavioral specification before this method can be used.

Who really determines the goals and objectives? Do they include all important outcomes? Have all those affected by the program agreed upon these particular goals or objectives? Who has determined that a particular criterion level is more defensible than alternatives? On what evidence? These and other questions must be addressed if an objectives-oriented approach is to be defensible.

Overemphasizing the testing components of this evaluation approach can prove dangerous. "Teaching for the test" is only human when a teacher's performance is evaluated by how well students do'on standardized or statewide assessment tests. Madaus (1983) describes how competency testing invariably turns into *,ninimunr* competency testing when expectations for achievement become bounded by test content. Such narrowing of educational purposes is a negative, albeit unintentional, consequence this approach may have had.

We should not leave our discussion of limitations of objectives—oriented evaluation without noting that Scriven's perception of its limitations led him to develop his now widely known proposals for goal-free evaluation (Scriven, 1972). Although intentionally the opposite of objectives-oriented approaches, it seems logical to discuss this proposal here.

Goal-Free Evaluation

The rationale for goal—free evaluation can be summarized as follows: First, educational goals should not be taken as given; like anything else, they should be evaluated. Further, goals are generally little more than rhetoric and seldom reveal the real objectives of the project or changes in intent. In addition, many important program outcomes do not fall in the category of goals or objectives anyway (for

example, establishing a new vocational education center will create additional jobs—a desirable outcome—but never an explicit goal of the center). Scriven believes the most important function of goal—free evaluation, however, is to reduce bias and increase objectivity. In objectives—oriented evaluation, an evaluator is told the goals of the project and is therefore immediately limited in her perceptions—the goals act like blinders, causing her to miss important outcomes not directly related to those goals.

For example, suppose an evaluator is told that the goals of a dropout rehabilitation program arc to (1) bring dropouts back into school, (2) train them in productive vocations, and (3) place them in stable jobs. She may spend all her time designing and applyirg measures to look at such things as how many dropouts have been recruited back into school, how many have been placed and remain placed in paying jobs, and so forth. These are worthwhile goals, and the program may be successful on all these counts. But what about the fact that the crime rate of other (nondropout) children in the high school has tripled since the dropouts were brought back into the school? Indeed, a hidden curriculum seems to have sprung up: stripping cars. This negative side effect is much more likely to be picked up by the goal—free evaluator than by the objectives—oriented evaluator working behind her built-in blinders.

The following are major characteristics of goal—free evaluation:

- The evaluator purposefully avoids becoming aware of the program goals.
- Predetermined goals are not permitted to narrow the focus of the evaluation study.
- Goal-free evaluation focuses on *actual* outcomes rather than intended program outcomes.
- The goal-free evaluator has minimal contact with the program manager and staff.
- Goal—free evaluation increases the likelihood that unanticipated side effects will be noted.

It might be helpful to point out that objectives-oriented and goal—free evaluation are not mutually exclusive. Indeed, they supplement one another. The internal staff evaluator of necessity conducts a goal—directed evaluation. She can hardly hope to avoid knowing the goals of the program, and it would be unwise to ignore them even if she could. Program managers obviously need to know how well the program is meeting its goals, and the internal evaluator uses goal—directed evaluation to provide such administrators with that information. At the same time, it is important to know how others judge the program, not only on the basis of how well it does what it is *supposed* to do but also on the basis of what it ***does*** in all areas, on *all* its outcomes, intended or not. This is the task for the external goal—free evaluator who knows nothing of the program goals. Thus,.goal-directed evaluati.pn and goal-free evaluation can work well together. And while the major share of a program's evaluation resources should not go to goal—free evaluation, it is tragic when all resources go to goal-directed evaluation on a program where the goals do not even begin to include the important educational outcomes.

APPLICATION EXERCISE

1. Mrs. Jackson is a member of the faculty of Greenlawn Middle School in Antioch, Ohio. Although students are still grouped by grades, within each grade a team of teachers cooperates to develop lessons that are interdisciplinary. Individual members of the team have been assigned responsibility for the areas of English. mathematics, science, and social studies. Mrs. Jackson has decided to evaluate her area of responsibility—the seventh—grade English section of the instructional program. Her evaluation tentatively includes:

a. Administration of a standardized English achievement test in September and June. She plans to compare the means of the pre- and posttest groups **with** national **norms for the tests.**

b. Monthly interviews **of** a 10 percent sample of her class to assess student reaction to the English portion of the instructional program.

c. Complete record keeping of students' progress so assessment may be made **of** their eighth-grade performance.

d. Observation by an outside observer twice a month, using a scale she has devised to record pupil interaction during class discussions.

e. Comparison of the performance of Mrs. Jackson's seventh-grade class on **the** standardized tests with the performance of the seventh grade at Martindale Junior High School, a traditional junior high.

Using what you have just learned about Tyler's approach to evaluation, how Hammond's cube can be used in evaluation, and how Provus's Discrepancy Evaluation Model works, advise Mrs. Jackson on her evaluation design. What questions should she be addressing? How could she organize her evaluation? How might she change her design to make it better?

SUGGESTED READINGS

BLOOM, B. S., **HASTINGS.** J. T., & I\4ADAUS, **G.** F. (1971). *J—jandboc'k* **OH** *froitit'' Ofll uflr,na— tine ei'aluario,g of srzidc,,t learnint.* New York: McGraw—Hill.

G00DLAD. **J.** (l979. *What schools are for.* Bloomington. IN: Phi Delta Kappa Educational Foundation.

HAMMOND, R. L. (1973). Evaluation at the local level. In B. R. W0WFUEN & j. R. S.NoERs. *Edizca,iv,,af rt',thiatio,,: Thcry and Prwtice.* Belmont, CA: Wadsworth.

METFESSEL N. S.. & **MiCHAEL,** W. B. (1967). A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. *Educational and Psychological Measure,nent, 27,* 931—943. Also in B. R. **W0RTHEN** & J. R. SANDERs (1973). *Educational Evaluation: Theory and Practice.* Belmont, CA: Wadsworth.

NOWAKOWSKI. J., **BUNDA,** M. A., **WORKING.** R., **BERNACKI, G..** & **HARRINCTON.** P. (1985).
A handbook of educational variables. Boston: Kluwcr—Nijhofl.

POPFIAM, W.J. (1975). *Educationa! evaluation.* Englewood Cliffs. NJ: Prentice—Hall.

Pitovus. M. (1971). *Dsscrepmcy evaluation.* Berkeley, CA: McCutchan.

S.irrn, E. R.. & **TYLIR.** R. W. (1942). *.4ppraisinc. and rrcordinç studcnt proc'rcs.* New. York:
Harper & Row.

DEVELOPERS OF THE MANAGEMENT-ORIENTED EVALUATION APPROACH AND THEIR CONTRIBUTIONS

The most important contributions to a management-oriented approach to eva1ua- tion in education have been made by Stuffiebeam and Alkin. in the mid-1960s, both *recognized* the shortcomings of available evaluation approaches. Working to expand and systematize thinking about administrative studies and educa-. tional decision making, they built upon concepts only hinted at in the earlier work of Bernard, Mann, Harris, and Washburne. During the 1960s and 1970s, they also drew from management theory (for example, Braybrooke & Lindblom, 1963). Both Stufflebeam and Alkin make the decision(s) ofprogram managers the pivotal organizer for the evaluation. Program objectives are not the main concern, In the models proposed by both theorists, the evaluator, working closely with administrator(s), identifies the decisions the administrator must make and then collects sufficient information about the relative advantages and disadvantages of each decision alternative to allow a fair judgment based on specified criteria. The success of the evaluation rests on the quality of teamwork between evaluators and decision—makers.

The CIPP Evaluation Model

Stufflebeam (1969, 1971, 1983; Stufflebeam & Shinktield, 1985) has been an influential proponent of a decision-oriented evaluation approach structured to help administrators make good decisions. He views evaluation as "the process of delineating, obtaining, and providing useful information for judging decision alternatives" (Stufilebeam, 1973a, p. 129). He developed an evaluation framework to serve managers and administrators facing four different kinds of educational decisions:

1. *Context evaluation,* to serve *planning decisions.* Determining what needs are to be addressed in an educational program helps in defining objectives for the program -
2. *Input evaluation,* to serve *structurinc decisions.* Determining what resources are available, what alternative strategies for the program should be considered, and what plan seems to have the best potential for meeting needs facilitates design of program procedures.
3. *Process evaluation,* to serve *implementing decisions.* How well is the plan being implemented? What barriers threaten its success? What revisions are needed? Once these questions arc answered, procedures can be monitored, controlled, and refined.
4. *Product evaluation,* to serve *recycling decisions.* What results were obtained? How well were needs reduced? What should be done with the program after it has run its course? These questions are important in judging program attainments.

The I9rst letter of each type of evaluation—context, input, process, and product—have been used to form the acronym "CIPP," by which Stufflebeam's evaluation model is best known. Table 6.1 summarizes the main features of the four types of evaluation, as proposed by Stufflcbearn and Shinkfield (1985, pp. 170—171).

As a logical structure for designing each type of evaluation, Stufflebeam proposed that evaluators follow these steps:

A. *Focusinç the Evaluation*

1. Identify the major level(s) of decision making to be served; fbr example, local, state, or national.
2. For each level of decision making, project the decision situations to be served and describe each one in terms of its locus, focus, criticality, timing, and composition of alternatives.
3. Detine criteria for each decision situation by specifying variables for measurement and standards for use in the judgment of alternatives.
4. Define policies within which the evaluator must operate.

B. *Coll&rion of Information*

1. Specify the source of the information to be collected.
2. Specify the instruments and methods for collecting the needed information.
3. Specify the sampling procedure to be employed.
4. Specify the conditions and schedulç, for information collection.

C. *Orcan ization o Information*

1. Provide a format for the information that is to be collected.
2. Designate a means for performing the analysis.

D. *Analysis of Information*

1. Select the analytical procedures to be employed.
2. Designate a means for performing the analysis.

E. *Reporting of Information*

1. Define the audiences for the evaluation reports.
2. Specify means for providing information to the audiences.
3. Specify the format for evaluation reports and/or reporting sessions.
4. Schedule the reporting of information.

F. *Administration of the Evaluation*

1. Summarize the evaluation schedule.
2. Define staff and resource requirements and plans for meeting these requirements.
3. Specify means for meeting policy requirements for conduct of the evaluation.
4. Evaluate the potential of the evaluation design for providing information that is valid, reliable, credible, timely, and pervasive.
5. Specify and schedule means for periodic updating of the evaluation design.
6. Provide a budget for the total evaluation program.

(Stufflebeam, 1973b, p. 144)

TABLE 6.1
Four Types of Evaluation

	Context Evaluation	*Input Evaluation*	*Process Evaluation*	*Product Evaluation*
Objective	To define the institutional context, to identify the target population and assess their needs, to identify opportunities for addressing the needs, to diagnose *problems* underlying the *needs*, and to judge whether proposed objectives are sufficiently responsive to the assessed needs.	To identify and assess *system capabilities*, alternative program *strategies*, procedural *designs* for implementing the strategies, budgets, and schedules.	To identify or predict in process, *defects* in the procedural design or its implementation, to provide information for the preprogrammed decisions, and to record and judge procedural events and activities.	To collect descriptions and judgments of outcomes and to relate them to objectives and to context, input, and process information, and to interpret their worth and merit.
Method	By using such methods as system analysis, survey, document review, hearings, interviews, diagnostic tests, and the Delphi technique.	By inventorying and analyzing available human and material resources, solution strategies, and procedural designs for relevance, feasibility and economy. And by using such methods as literature search, visits to exemplary programs, advocate teams, and pilot trials.	By monitoring the activity's potential procedural barriers and remaining alert to unanticipated ones, by obtaining specified information for programmed decisions, by describing the actual process, and by continually interacting with, and observing the activities of project staff.	By defining operationally and measuring outcome criteria, by collecting judgments of outcomes from stakeholders, and by performing both qualitative and quantitative analyses.
Relation to decision making in the change process	For deciding upon the *setting* to be served, the *goals* associated with meeting needs or using opportunities, and the *objectives* associated with solving problems; that is, for *planning* needed changes. And to provide a basis for judging outcomes.	For selecting *sources of support*, solution *strategies*, and procedural *designs*; that is, for *structuring* change activities. And to provide a basis for judging implementation.	For *implementing and refining the program design and procedure*; that is, for effecting *process control*. And to provide a log of the actual process for later use in interpreting outcomes.	For deciding to *continue*, *terminate*, *modify*, or *refocus* a change activity. And to present a clear record of effects (intended and unintended, positive and negative).

Source: Stufflebeam and Shinkfield, 1985, pp. 170–171

The UCLA Evaluation Model

While he was director of the Center for the Study of Evaluation at UCLA, Alkin (1969) developed an evaluation framework that paralleled closely some aspects of the CIPP model. Alkin defined evaluation as "the process of ascertaining the decision areas of concern, selecting appropriate information, and collecting and analyzing information in order to report summary data useful to decision—makers in selecting among alternatives" (Alkin, 1969, p. 2). Alkin's model included the following five types of evaluation:

1. *Systems assessment7* to provide information about the state of the system. (Very similar to context evaluation in the CIPP model.)

2. *Proran, plannitiç,* to assist in the selection of particular programs likely to be effective in meeting specific educational needs. (Very similar to input evaluation.)

3. *Proraiii implementation,* **to** provide information about whether a program was introduced to the appropriate group in the manner intended.

4. Pro,irani improvement, to provide information about how a program is functioning, whether interim objectives are being achieved, and whether unanticipated outcomes are appearing. (Similar to process evaluation.)

3. *Pro'ra,n* certification, to provide information about the value of the program and its potential for use elsewhere. (Very similar to product evaluation.)

l3oth rhe CIPP and UCLA frameworks for evaluation appear to be linear and sequential, but the developers have stressed that such is not the case. For example, the evaluator would not have to complete an input evaluation or a systems assessment in order to undertake one of the other types of evaluation listed in the framework. Often evaluators may undertake "retrospective" eyaluations (such as a context evaluation or a systems assessment) in preparation for a process or program improvement evaluation study, believing this evaluation approach is cumulative, linear, and sequential; however, such steps are not always necessary. A process evaluation can be done without having completed context or input evaluation studies. At other times, the evaluator may cycle into another type of evaluation if some decisions *suggest* that earlier decisions should be reviewed. Such is the nature of management-oriented evaluation.

More recent work on the CIPP model has produced working guides for types of evaluation included in that framework. For example, Stufflebeam (1977) advanced the procedure for conducting a *context* evaluation with his guidelines for designing a needs assessment for an educational program or activity.

A guide for use in *input* evaluation was developed by Reinhard (1972). The input evaluation approach that she developed is called the *advocate team technique.* It is used when acceptable alternatives for designing a new program are not available or obvious. The technique creates alternative new designs that are then evaluated and selected, adapted, or combined to create the most viable alternative design for a new program. This technique has been used successfully by the federal government (Reinhard, 1972) and by school districts (Sanders, 1982) to generate options and guide the final design of educarior2al programs.

Procedures proposed by Cronbach (1963) provided useful suggestions for the conduct of *process* evaluation.

Techniques discussed in Chapter 18 of this book (most notably goal-free evaluation and comparative experimental designs) provide information useful in conducting *product* evaluations, as do approaches to evaluation discussed later in Chapter 7.

Other Management-Oriented Evaluation Approaches

In Chapter **5,** Provus's Discrepancy Evaluation Model was described as an objectives—oriented evaluation model. Some aspects of that model are also directed toward serving the information needs of educational program managers. It is systems—oriented, focusing on input, process, and output at each of five stages of evaluation: program definition, program installation, program process, program products, and cost-benefit analysis. Even cursory scrutiny of these five types of evaluation reveal close parallels to the CIPP and UCLA evaluation models with respect to their sensitivity to the various decisions managers need to make at each stage of program development.

The utilization-focused evaluation approach of Patton (1978) in one respect could also be viewed as a decision—making approach. He stressed that the process of identifying and organizing relevant decision—makers and information—users is the first step in evaluation. In his view, the use of evaluation findings requires that decision-makers determine what information is needed by various people and arrange for that information to be collected and provided to those persons.

The systems analysis approach has also been suggested by some to be an evaluation approach (for example, House, 1980; Rivlin. 1971; Rossi, Freeman, & Wright, 1979) and if we agreed we would place it with others covered in this chapter. However, we do not consider systems studies to be evaluation because of their narrow research focus on establishing causal links between a few preselected variables and on cost analyses. We consider such studies to be good examples of social research rather than evaluation.

HOW THE MANAGEMENT-ORIENTED EVALUATION APPROACH HAS BEEN USED

The CIPP model has been used in school districts and state and federal government agencies. The Dallas, Texas, Independent School District, for example, established an evaluation office organized around the four types of evaluation in the model. All evaluation activities in that district fall into one or more of these categories. The Saginaw, Michigan, public schools also strucrured their evaluation work by using the CIP1 model, as did the Lansing, Michigan, public schools, and evaluations systems in the Cincinnati and Columbus public school systems in Ohio.

The management—oriented approach to evaluation has guided educators through program planning, operation, and review. Program staff have found this approach a useful guide to program improvement.

This evaluation approach has also been used for accountability purposes. It

provides a record-keeping framework that facilitates public review of educational needs, objectives, plans, activities, and outcomes. School administrators and school boards have found this approach useful in meeting public demands for information.

Stufflebeam and Shinkfield (1985) described these two uses of the CIPP model as shown in Figure 6.1:

	Decision Making (Formative Orientation)	*Accountability (Summative Orientation)*
Context	Guidance for choice of objectives and assignment of priorities.	Record of objectives and bases for their choice along with a record of needs, opportunities, and problems.
Input	Guidance for choice of program strategy. Input for specification of procedural design.	Record of chosen strategy and design and reasons for their choice over other alternatives.
Process	Guidance for implementation.	Record of the actual process.
Product	Guidance for termination, continuation, modification, or installation.	Record of attainments and recycling decisions.

FIGURE 6.1 The Relevance of Four Evaluation Types to Decision Making and Accountability

Source: Stufflebeam and Shinkfield, 1985, p. 164.

STRENGTHS AND LIMITATIONS OF THE MANAGEMENT-ORIENTED EVALUATION APPROACH

This approach has proved appealing to many evaluators and program managers, particularly those at home with the rational and orderly systems approach, to which it is clearly related. Perhaps its greatest strength is that it gives focus to the evaluation. Experienced evaluators know how tempting it is simply to cast a wide net, collecting an enormous amount of information, only later to discard much of it because it is not directly relevant to the key issues or questions the evaluation must address. Deciding precisely what information to collect is essential. Focusing on informational needs and pending decisions of managers limits the range of relevant data and brings the evaluation into sharp focus. As House put it (using the term "decision making approach" to describe the CIPP model and related approaches),

> the decision making approach provides a valuable insight into evaluation. It stresses the importance of the utility of information. Evaluation information is meant to be used. Connecting evaluation to decision-making underlines the purpose of evaluation. It is also practically useful to shape an evaluation in reference to actual decision-making considerations. Even if one cannot define precisely the decision alternatives, one can eliminate a number of lines of inquiry as being irrelevant. (House, 1980, p. 232)

The management-oriented approach to evaluation was instrumental in showing evaluators and educators that they need not wait until an activity or program has run its course before evaluating it. In fact, educators can begin evaluating even when ideas for programs are first discussed. Because of lost opportunities and heavy resource investment, evaluation is generally least effective at the end of a developing program. Of course, educators have found that it's never too late to begin evaluating, even if a program has been in place for years. The decisions are simply different.

The management-oriented evaluation approach is probably the preferred choice in the eyes of most school administrators and boards. This is hardly surprising given the emphasis this approach places on information for decision-makers. By attending directly to the informational needs of people who are to use the evaluation, this approach addressed one of the biggest criticisms of evaluation in the 1960s: that it did not provide useful information.

The CIPP model, in particular, is a useful and simple heuristic tool that helps the evaluator generate potentially important questions to be addressed in an evaluation. For each of the four types of evaluation (CIPP), the evaluator can identify a number of questions about an educational undertaking. The model and the questions it generates also make the evaluation easy to explain to lay audiences.

The management-oriented approach to evaluation supports evaluation of every component of an educational program as it operates, grows. or changes. It stresses the timely use of feedback by decision—makers so that education is not left to flounder or proceed unatTected by updated knowledge about needs, resources, new developments in education, the realities of day—to—day operations. or the consequences of providing education in any given way.

A potential weakness of this approach is the evaluator's occasional inability to respond to questions or issues that may be significant—even critical—but that clash with or at least do not match the concerns and questions of the decision—maker who essentially controls the evaluation.

House also issued a warning when he asked,

Why should the decision—maker, who is usually identified as the program administrator, be given so much pretrence? I)oes this not put the evaluator at the service of top
management and make the evaluator the "hired gun" of the program establishment? Does this not make the evaluation potentially unfair and even undemocratic? The answer is that these are potential weaknesses of the decision—making approach. (House. 1980,
p. 231)

To build on House's point, we might want to consider the policy uses of evaluations by what Cronbach and others (1980) called the *policy-shaping corn,nuni(y*. The policy-shaping community includes: (1) public servants, such as responsible officials at the policy and program levels and the actual operating personnel; and (2) the public, consisting not only of constituents, but also influential persons such as commentators, academic social scientists, philosophers, gadflies, and even novelists or dramatists. Few policy studies have been found to have *I* a direct effect on the policy—shaping community, but evaluations -can and do

influenec these audiences over time. Policy, as a reflection of public values, may be seen as a never-ending aspect of education that continues to be molded or revised as issues, reforms, social causes, and social values change or come to the forefront of attention. We need to remember, as Cronbach has noted, that one important role of the evaluator is to illuminate, not to dictate, the decision. Helping clients to understand the complexity of issues, not to give simple answers to narrow questions, is a role of evaluation.

Another limitation is that, if followed in its entirety, the management—oriented approach can result in costly and complex evaluations. If priorities arc not carefully set and followed, the many questions to be addressed using a management— oriented approach can all clamor for attention, leading to an evaluation system as large as the program itself and diverting resources from program activities. In planning evaluation procedures, management-oriented evaluators need to consider the resources and time available. If the management—oriented approach requires more time or resources than are available, another approach may have to be considered.

As a case in point, consider the classroom teacher who has to make decisions about next week's lesson plans. Because of his time limitations, and the limited information that is readily available to him, this teacher may be able to use only the C[PP or UCLA models informally, as an armchair aid. As with any approach, the management-oriented evaluator needs to be realistic about what work is possible.and not to promise more than can be delivered.

Finally, this evaluation approach assumes that the important decisions can be clearly identified in advance, that clear decision alternatives can be specified, and that the decisions to be served remain reasonably stable while the evaluation is being done. All of these assumptions about the orderliness and predictability of the decision-making process are suspect and frequently unwarranted. Frequent adjustments may be needed in the original evaluation plan if this approach is to work vell.

APPLICATION EXERCISE

A public school system successfully demonstrated its need for federal support for an elementary compensatory education program. They received a $500,000 grant to be spent over a period of three years from july 1, 1986 to june 30, 1989. On March 15, 1986, the Superintendent convened a meeting of the Assistant Superintendent of Elementary Instruction and 30 principals of elementary schools eligible to participate in the proposed program. It was their decision that a thorough evaluation of the reading and mathematics programs in these schools should be completed by September 30, 1986 to identify needs. Alternative strategies for solving needs would then be evaluated and a program would be chosen for the elementary compensatory education project. They also decided to establish an evaluation team that would be responsible for:

1. Conducting the evaluation of the reading and mathematics programs of the eligible schools

2. Evaluating alternative programs to meet the needs of the 30 schools
3. Continually monitoring the program, which would be implemented starting in 1986
4. Collecting information to be reported annually (on June 30 for each year of the grant) to the United States Department of Education.

Using what you have just learned about management-oriented evaluation approaches, advise the evaluation team members about how they should proceed (assuming that it is now March 1986). Be as detailed in your planning as you can be.

SUGGESTED READINGS

Aucn.i, M. C. (1969). Evaluation theory development. *Epaluation Comment, 2,* 2—7.

Also excerpted in B. R. **WORTHEN** & **J.** R. **SANDERS** (1973), *Educational evaluation: Theory and Practice*. Belmont, CA: Wadsworth.

STUFPLEBEAM, D. L. (1983). The CIPP model for program evaluation. In G. F. **MADAUS,** M. SCRIVEN, & D. L. STUFELEBEAM (Eds.), *Evaluation models*. Boston: Kluwer-Nijhoff.

STUFFLEBEAM, D. L., FOLEY, W.J., GEPHART, W.J, GUBA, E. G.. HAMMOND, R. L., MER1UMAN, H. 0., & **PROVUS,** M. M. (1971). *Educational evaluation and decision making*.
Itasca, IL: F.E. **PEACOCK.**

STUFFLEBEAM, D. L., & **SHINKFIELD,** A.J. (1985). *Systematic evaluation*. Boston: KluwerNijhof[

rating products and product evaluation reports are two typical outgrowths of this approach.

The consumer-oriented approach to evaluation is predominantly a summative evaluation approach. Developers of educational products have come to realize, however, that using the checklists and criteria of the consumer advocate while the product is being created is the best way to prepare for subsequent public scrutiny. Thus, the checklists and criteria proposed by "watchdog" agencies have become tools for formative evaluation of products still being developed.

DEVELOPERS OF THE CONSUMER-ORIENTED EVALUATION APPROACH AND THEIR CONTRIBUTIONS

The importance of consumer-oriented evaluation seems to have been first recognized during the mid- and late 1960s as new curriculum packages and other educational products began to flood the market. Prior to the 1960s, most materials available to educators were textbooks. With the influx of funds earmarked for product development and federal purchases, however, the marketplace swelled.

Scriven's Concerns and Checklists

Scriven (1967) made a major contribution to this approach with his distinction between formative and summative evaluation. The summative role of evaluation. he said, "[enablesj administrators to decide whether the entire finished curriculum, refined by use of the evaluation process in its. . . [formativel role, represents a sufficiently significant advance on the available alternatives to justify the expense of adoption by a school system" (Scriven, 1967, pp. 41—42). Criteria that Scriven suggested for evaluating any product included the following:

- Evidence of achievement of important educational objectives
- Evidence of achievement of important noneducational objectives (for example. social objectives)
- Follow-up results
- Secondary and unintended effects, such as effects on the teacher, the teacher's colleagues, other students, administrators, parents, the school, the taxpayer, and other incidental positive or negative effects
- Range of utility (that is, for whom will it be useful)
- Moral considerations (unjust uses of punishment or controversial content)
- Costs.

Later, Scrivcn (1974b) published a product checklist that expanded his earlier criteria. This new product checklist was the result of reviews commissioned by the federal government, focusing on educational products developed by federally sponsored research and development centers and regional laboratories, It was used in the examination of over 90 educational products, most of which underwent many revisions during that review. Scriven stressed that the items in this checklist were *necessitata,* not *desiderata*. They included the following:

1. *Need*. Number affected, social significance, absence of substitutes, multi—plicative effects, evidence of need.
2. *Market*. Dissemination plan, size, and importance of potential markets.
3. *Performance—True Field Trials*. Evidence of effectiveness of final version, with typical users, with typical aid, in typical settings, within a typical time frame,
4. *Performance—True Consumer*. Tests run with all relevant "Consumers," such as students, teachers, principals, school district staff, state and federal officials, Congress, and taxpayers.
5. *Performance—Critical Comparisons*. Comparative data provided on important competitors such as no—treatment groups, existing competitors, projected competitors, created competitors, and hypothesized competitors.
6. *Performance—Long Term*. Evidence of effects at pertinent times is reported, such as a week to a month after use of the product, a month to a year later, a year to a few years later, and over critical career stages.
7. *Performance—Side Effects*. Evidence of independent study or search for unintended outcomes during, immediately following, and over the long- term product use.
8. *Performance—Process*. Evidence of product use provided to verify product descriptions, causal claims, and the morality of product use.
9. *Performance—Causation*. Evidence of product effectiveness provided through randomized experimental study or through defensible quasi-experimental, ex post facto, or correlational studies.
10. *Performance—Statistical Signflcance*. Statistical evidence of product effectiveness to make use of appropriate analysis techniques, significance levels, and interpretations.
11. *Performance—Educational Significance*. Educational significance is demon strate through independent judgments, expert judgments, judgments based on item analysis and raw scores of tests, side effects, long-term effects and comparative gains, and educationally sound use.
12. *Cost-effectiveness*. A comprehensive cost analysis, including expert judg men of costs, independent judgment of costs, and comparison to costs for competitors.
13. *Extended Support*. Plans for postmarketing data collection and improvement, in-service training, updating of aids, and study of new uses and user data.

These are stringent standards, to be sure, but defensible and important—although few textbooks or curriculum packages now on the market would satisfy all of them. Perhaps no one educational product will ever be judged successful on all these criteria, but producers' efforts to meet these standards would have a marked effect on improving the efforts of educational developers.

Scriven continues to be the most avid and articulate advocate of the consumer—oriented evaluation approach, although he is not blind to weaknesses in some of its applications, as noted in the following observation:

We should add a word about what may seem to be the most obvious of all models for a consumerist ideologue, namely *Consumer Reports* product evaluations. While these serve

as a good enough model to demonstrate failures in most of the alternatives more widely accepted in program evaluation, especially educational program evaluation, it must not be thought that the present author regards them as flawless. I have elsewhere said
something about factual and logical errors and separatist bias in *Consumer Reports* ("Product Evaluation" in N. Smith, ed., *New Models of Program Evaluation,* Sage, 1981). Although *Consumer Reports* is not as good as it was and it has now accumulated even
more years across which the separatist/managerial crime of refusal to discuss its methodologies and errors in an explicit and nondefensive way has been exacerbated many times, and although there are now other consumer magazines which do considerably
better work than *Consumer Reports* in particular fields, *Consumer Reports* is still a very good model for most types of product evaluation. (Scriven, 1984, -p. 75)

Other Checklists and Product Analysis Systems

In the mid—1960s, Komoski was a leader in establishing the Educational Products Information Exchange (EPIE) as an independent product review service modeled after the Consumers Union. Through its newsletter *(EPIE Forum)* and published reports,28 EPIE has provided much needed evaluative information to state departments of education and school districts that subscribe to its service. EPIE checklists and curriculum analysis guides have also been valuable tools for the educational consumer. In addition, EPIE collects and disseminates information being used in school settings.

Likewise, checklists developed by Morrisett and Stevens (1967), Tyler and Klein (1967), and Eash (1970) have been useful aids to educators responsible for compiling information for use in selecting educational products. The Curriculum Materials Analysis System (CMAS) developed by Morrisett and Stevens, for example, includes the following guidelines for product analysis:29

1. *Describe the characteristics of the product:* Media, materials, time needed, style, costs, availability, available performance data, subject matter and content, dominant characteristics of curriculum forms.
2. *Analyze its rationale and objectives:* Describe and evaluate rationale, general objectives, specific objectives, behavioral objectives -
3. *Consider antecedent conditions in using this product:* Pupil characteristics, teacher capabilities and requirements, community and school characteristics, existing curriculum and curriculum organization (vertical and horizontal).
4. *Consider its content:* The cognitive structure, skills to be taught, affective content.
5. *Consider the instructional theory and eaching strategies used in this product:* Appropriateness of teaching strategies, forms, modes, or transactions.
6. *Form overall judgments:* Be sure to consider other descriptive data, reported experiences with the product, pilot tryouts, and outside recommendations.

A variety of product evaluation checklists have been developed and used over the past several years by individual evaluators and agencies. Many serve as valuable guides from which one might develop a checklist tailored for one's own situation. The following checklists developed by the Florida Department of Education (1980) and Patterson (undated) provide good examples of concise review forms useful for heuristic purposes.

FLORIDA DEPARTMENT OF EDUCATION

Checklist for Reviewing and Selecting Materials[30]

LEVEL I

Equipment and Storage Considerations

	yes	no	don't know	N/A*
If equipment is needed to use the materials, is the equipment available and accessible?				
Is space available in the classroom to store the materials?				
Is work space available in the classroom so that materials can be used properly?				

Materials-Use Considerations

	yes	no	don't know	N/A*
Are teachers' guides included in the materials?				
If the materials require innovative or unusual strategies, are there indications that the strategies used in the materials can easily be integrated into your present system, without having long-term negative effects on student learning and classroom management?				
Have materials been field-tested in similar situations and found to be successful?				

Cost Considerations

	yes	no	don't know	N/A*
Is there enough money in the budget?				
Will the materials be used often enough to justify the cost?				
Can parts of the materials be purchased separately from the overall package?				
Can you afford to continue using the materials?				
Are materials well made and likely to withstand repeated use?				

*Not applicable

LEVEL II

Content Considerations

	yes	no	don't know	N/A*
Does the content match course objectives?				
Is the content appropriate?				

Functions-of-Instruction Considerations

	yes	no	don't know	N/A*
Are the materials likely to gain and maintain the learner's attention?				
Are the materials likely to inform the learner of the objective?				
Do the materials provide for recall of relevant learning?				

Do the materials present learning content in appropriate ways?				
Do the materials provide learning guidance?				
Do the materials provide an opportunity to practice?				
Is feedback on the performance provided within the materials?				
Are there provisions within the materials for assessing the learner's performance?				
Do the materials provide for retention and transfer?				
Instructional Strategies Considerations				
Are materials practical and easy to use?				
Stereotyping and Bias Considerations				
Do materials avoid stereotypes and bias?				

*Not applicable

Source. Center for Instructional Development and Services, Florida State University

HOW THE CONSUMER-ORIENTED EVALUATION APPROACH HAS BEEN USED

As mentioned previously, the consumer-oriented approach to evaluation has been used extensively by government agencies and independent educational consumer advocates such as EPIE to make information available on hundreds of products.

The Joint Dissemination Review Panel (JDRP) in the United States Department of Education has established a classic example of a consumer-oriented evaluation system by setting standards (Tallmadge, 1977) for new educational programs that it will recommend for adoption. Those that pass the JDRP review are approved through the National Dissemination Network for dissemination to school systems throughout the United States.

A role for state departments of education or other educational service agencies, such as intermediate school districts, in using this approach has been discussed by several states in recent years. Rather than evaluating products themselves, and then disseminating their findings to school districts, or merely listing available products, states and service agencies have discussed using standard forms to compile and then disseminate evaluation information about any new product. Guidelines developed for this purpose (Sanders, 1974) address four aspects of a product: its educational processes, content, transportability, and effectiveness. In each case the central concern has been, "What does one need to know about a product before deciding whether to adapt or install it?" Questions posed within each category include the following:

Process Information

1. What is the nature and frequency of interactions among (combinations of) students/teachers/administrators/relevant others? Have these interactions been evaluated?

Instructional Materials Review Form:

Marvin Patterson
Center for Studies in
Vocational Education
Florida State University

Title(s) _____

Author(s) _____

Publisher _____

Latest Copyright Date _____

- ☐ Retain for Committee Review
- ☐ Bibliography Only
- ☐ Reject (Comments:

Use the following code to rate materials:

\+ means yes or good quality — means no or poor quality
o means all right, but not of especially good quality NA means not applicable

Committee Members
1. Does the content cover a significant portion of the program competencies?
2. Is the content up-to-date?
3. Is the reading/math level appropriate for most students?
4. Are objectives, competencies, or tasks stated in the student materials.
5. Are tests included in the materials?
6. Are performance checklists included?
7. Are hands-on activities included?
8. How many outside materials are required? \+ means 0-1 materials o means 2-3 materials — means 4+ materials
9. Would you use these materials in your training program?
If the materials to this point appear to be a possible choice for selection, continue with your review. Stop if the materials appear to be too poor for further consideration.
10. Is a Teacher's Guide included which offers management suggestions for the materials?
11. Is the material presented in a logical sequence?
Quality Judgments. Use +, o, — to rate the quality of the products.
12. Quality of objectives, competencies, and/or tasks
13. Degree of match between learning activities and objectives.
14. Quality of test items and degree of match with objectives.
15. Quality of performance checklists and degree of match with objectives.
16. Quality of directions for how students are to proceed through the materials.
17. Quality of drawings, photographs, and/or other visuals.
18. Overall design of the learning activities for individualized instruction.
19. Emphasis on safety practices (when needed)
20. Degree of freedom from bias with respect to sex, race, national origin, age, religion, etc. (see provided guidelines).
21. Quality of management procedures for teachers (teacher's guides, etc.).
22. (Optional) List the career-map competencies covered by these materials.

Comments:

Source: Center for Instructional Development and Services, Florida State University

2. Is the teaching strategy to be employed described so that its appropriateness ***can*** be determined? Has the strategy been evaluated?
3. Is the instructional schedule required by the program or product described so that its feasibility can be determined? Has the schedule been evaluated?
4. Are the equipment and facilities required by the program or product described so that their feasibility can be determined? Have they been evaluated?
5. Are the budget and human resource requirements of the program or product listed so that their feasibility can be determined? Have the following require— ments been included?
a. Start-up and continuation budget requirements?
b. Administration/teaching staff/parent or advisory committees/student resource requirements?
c. In—service?
6. Is evaluation an integral part of the (a) development, and (b) implementation of the program or product?

Content Information

1. Is the existence or lack of basic program or product elements, such as the following, noted?
a. Clearly stated objectives?
b. Sufficient directions?
c. Other materials required?
d. Prerequisite knowledge/attitudes required?
e. Fit with disciplinary knowledge base and existing curriculum and sequence of a school?
2. Have these elements been evaluated?
3. Has the content of the program or product been evaluated by recognized content specialists?
4. Is there sufficient information about the program or product rationale and philosophy to permit a decision about whether it is within the realm of school responsibility or consistent with a district's philosophy? (School responsibility has traditionally encompassed such areas as intellectual development, personal development, citizenship. social development, educational and personal adjustment, physical growth and development.)
5. Do the objectives of the program or product have cogent rationales? Pre they compatible with the school philosophy and the values of the community?

Transportability Information

Referring to the Bracht and Glass (1968) treatment of external validity, three elements relating to transportability appear critical: (1) geography and setting, (2) people, and (3) time.
I. What information is available regarding classroom/school/community settings in which the program or product has been used effectively; in terms of:

a. Size/numbers?
b. Organization?
c. Political factors?
d. Legal issues?
e. Facilities?
f. Wealth?
g. Occupational factors?
h. Geographical indices (for example, rural/urban)?
i. Cultural factors?
j. Public/nonpublic factors?
k. Philosophical issues?
2. What information is available concerning the teachers/students /administra—tors/relevant others with whom the program or product has been used effectively in relation to the following:
a. Age/grade?
b. Experience?
c. Entrance knowledge?
d. Expectations/preferences/interests?
e. Ethnic/cultural makeup?
3. What information is available regarding the time of year in which the program or product has been used effectively?
4. Does the program or product have special requirements in areas such as the following:
a. Training?
b. Organization or facilities?
c. Additional materials or equipment?
d. Research people/specialists?
Effectiveness **Information**
1. Has technically sound information about the effects of the program or product on its target audience been obtained using one of more of the following procedures:
a. Comparison to established base (pre.-post)?
b. Prediction (make success estimates using regression techniques)?
c. Comparison to local or state norms?
d. Comparison against objectives or predetermined standards?
e. Comparison of student and/or teacher behaviors to those for competing programs or products?
2. Is there evidence that the program or product has eliminated a documented need, such as
a. Cost savings?
b. Improved morale?
c. Faster learning rate?
3. Is immediate and follow—up effectiveness information available?

These questions could easily be cast into a checklist formed by using the following response categories:

YES NO IMPLIED NOT APPLICABLE

Information available:

Although not necessarily a product evaluation enterprise, the *Mental Measurement Yearbooks* (Mitchell, 1985) must be *mentioned* as a form of consumer-oriented evaluation. These yearbooks contain critical reviews of commercially available tests marketed in the United States and several other English—speaking countries. As educational products, these tests deserve the same scrutiny any product receives, and the *Mental Measurement Yearbooks* have provided a valuable service in this regard to educators.

STRENGTHS AND LIMITATIONS OF THE CONSUMER-ORIENTED EVALUATION APPROACH

Developers of the consumer-oriented approach to evaluation have provided a valuable service to educators in two ways: (I) they have made available evaluations of educational products as a service to educators who ma not have the time or information to do the job thoroughly; and (2) they have advanced the know!edge ofrducators about the criteria most appropriate to use in selecting educational products. The checklists that have evolved from years of consumer-oriented product evaluations are useful and simple evaluation tools.

Consumers have become more aware of the sales ploys of the education industry, thanks to the efforts of consumer—oriented evaluators. Educators are (or should be) less vulnerable to sales tactics than they were 20 years ago. They are (or should be) more discriminating in the way they select products.

The educational product industry has a long way to go, however, in being responsive to educators' needs. Just ask a textbook sales rep for information about the performance or proven effectiveness of his product, and see what kind of information you get. Most of the time it will be either anecdotal testimony or scanty, poorly conceived product evaluation. Very seldom do corporations spend the time or money needed to acquire acceptable information about their products' performance. Educational consumers must insist on such information if the educational product industry is to take product evaluation seriously.

The consumer-oriented approach to evaluation is not without drawbacks (although they seem small compared to its benefits). It can increase the cost of educational products. The time and money invested in product testing will usually be passed on to the consumer. Moreover, the use of stringent standards in product development and purchase may suppress creativity because of the risk involved. There is a place for trial programs in schools before products are fully adopted. By

developing long-term plans for change, educators can give untested products a chance without consuming large portions of the budget. Cooperative trial programs *with* the educational product industry *may* be in the best interest of all.

Finally, the consumer—oriented approach to evaluation threatens local initiative development because teachers or curriculum specialists may become increasingly dependent on outside products and consumer services. Educators need to place the purchase of outside products in perspective so that they are not overly dependcnt on the availability of other people's work. We agree with those who contend that we need to be less concerned with developing and purchasing teacher-proof products and more concerned with supporting product-proof teachers—teachers who think for thems'elvcs and who take the initiative in addressing student needs.

APPLICATION EXERCISE

Teachers in each academic-department of a senior high school are given a choice of curricular programs to use in their classrooms. This policy is designed to take full advantage of the range of capabilities of individual faculty members. All faculty members are required to prepare an evaluation of their program and circulate these reports among the faculty. These evaluations are to be conducted in keeping with the evaluation guidelines established by the curriculum council. The guidelines require:

1. A statement from each teacher at end of year about the goals for his course and an assessment of the extent to which these goals were met using the selected program.
2. Submission of information from each reacher about the comparison of his pupils' performance and the performance of pupils using a different curricular program. -.
3. An outside assessment of the appropriateness of the selected program for the teacher's stated goals.
4. A comparison of student performance on standardized tests to national norms.
5. A complete list of test items used during the year. Results for each item to be reported.

Using what you have just learned about consumer-oriented evaluation, what changes in the evaluation process would you suggest to the curriculum council? How could faculty reports be structured so that other schools could benefit from their consumer reports?

SUGGESTED READINGS

SCRIVEN, M. (1974b). Standards for the evaluation of educational programs and products. In

G. D. Boiticit (Ed.), *Evaluating educational programs and products.* Englewood Cliffs, NJ:

Educational Technology Publications. Also in W.J. POPHAM (Ed.). (1974). *Evaluation in education.* Berkeley, CA: McCutchan.

TALLMADCE, G. K. (1977). *Ideabook: JDRP*. (ERIC DL 48329). Washington, DC: U.S. Government Printing Office.

Chapter 8
Expertise-Oriented Evaluation Approaches

Orienting Questions

1. What are the arguments for and against using professional judgment as the means for evaluating educational programs?
2. Under what conditions would accreditation, blue-ribbon panels, or connoisseurship be methods of choice for conducting an evaluation?
3. What criteria would you use to screen experts in order to select the best for an expertise-oriented evaluation?
4. What differences exist between formal and informal professional review systems?
5. What are some major strengths and limitations of the expertise-oriented evaluation approach?

The expertise-oriented approach to evaluation, probably the oldest and most widely used, depends primarily upon professional expertise to judge an educational institution, program, product, or activity. For example, the worth of a curriculum would be assessed by curriculum or subject-matter experts who would observe the curriculum in action, examine its content and underlying learning theory or, in some other way, glean sufficient information to render a considered judgment about its value.

Although subjective professional judgments are involved to some degree in all the evaluation approaches described thus far, this approach is decidedly different because of its direct, open reliance on subjective professional expertise as the primary evaluation strategy. Such expertise may be provided by the evaluator(s) or by someone else, depending on who offers most in the substance or procedures being evaluated.

Several specific evaluation processes are variants of this approach, including doctoral oral examinations, proposal review panels, professional reviews conducted by professional accreditation bodies, reviews of institutions or individuals by state or national licensing agencies, reviews of educators' performance for decisions concerning promotion or tenure, peer reviews of articles submitted to "refereed" professional journals, site visits of educational programs conducted at

the behest of the program's sponsor, reviews and recommendations of prestigious "blue-ribbon" panels, and even the critique offered by the ubiquitous expert who exists, at least one to every educational system, and whose *raison d'etre* is to serve in a self-appointed watchdog role.

To impose some order, we choose to organize and discuss these various manifestations of expertise-oriented evaluation within four categories: (1) formal professional review systems; (2) informal professional review systems; (3) ad hoc panel reviews; and (4) ad hoc individual reviews. Differences in these categories are shown in Table 8.1, along the following dimensions:

1. Is there an existing structure for operating the review?
2. Are published standards used as part of the review?
3. Are reviews scheduled at specified intervals?
4. Does the review include opinions of multiple experts?
5. Do results of the review have an impact on the status of whatever is reviewed?

To this we have added a fifth category, namely educational connoisseurship and criticism, to discuss an interesting expertise-oriented approach that does not fit neatly into the other categories or dimensions shown in Table 8.1.

Site visitation, frequently the mode for conducting expertise-oriented evaluations, is not itself an approach to evaluation; rather, it is a method that might be used not only here but also with other evaluation approaches. Site-visit methods and techniques are discussed in Chapter 18, along with other techniques often used by expertise-oriented evaluators.[31]

DEVELOPERS OF THE EXPERTISE-ORIENTED EVALUATION APPROACH AND THEIR CONTRIBUTIONS

It is hard to pinpoint the origins of this approach, for it has been long with us. It was formally used in education in the 1800s, when schools began to standardize college entrance requirements. Informally, it has been in use since the first time an

TABLE 8.1
Some Features of Four Types of Expertise-Oriented Evaluation Approaches

Type of Expertise-Oriented Evaluation Approach	*Existing Structure*	*Published Standards*	*Specified Schedule*	*Opinions of Multiple Experts*	*Status Affected by Results*
Formal Review System	Yes	Yes	Yes	Yes	Usually
Informal Review System	Yes	Rarely	Sometimes	Yes	Usually
Ad Hoc Panel Review	No	No	No	Yes	Sometimes
Ad Hoc Individual Review	No	No	No	No	Sometimes

individual to whom expertise was publicly accorded rendered a judgment about the quality of some educational endeavor (and history is mute on when that occurred). Several movements, and individuals have given impetus to the various types of expertise—oriented evaluations, as described below.

Formal Professional Review Systems

We would define a formal professional review system as one having (1) structure or organization established to conduct periodic reviews of educational endeavors; (2) published standards (and possibly instruments) for use in such reviews; (3) a prespecified schedule (for example, every five years) on which reviews will be conducted; (4) opinions of several experts combining to reach the overall judgments of value; and (5) an impact on the status of that which is reviewed, **depending on the outcome.**

Accreditation. To many, the most familiar formal professional review system is that of *accreditation,* **the process whereby an** organization grants approval of educational institutions. Beginning in the late 1800s, national and regional accreditation agencies gradually supplanted in the United States the borrowed Western European system of school inspections, and these became a potent force in education during the 1930s. Education was not alone in institutionalizing accreditation processes to determine and regulate the quality of its institutions. Parallel efforts were underway in other professions, including medicine and law, as concern over quality led to wide scale acceptance of professionals judging the efforts of fellow professionals. Perhaps the most memorable example is Flexner's (1910) examination of medical schools in the United States and Canada in the early 1900s, which led to the closing of numerous schools he cited as inferior. As Floden (1983) has noted, Flexner's study was not accreditation in the strict sense, because medical schools did not participate voluntarily, but it certainly qualified as accreditation in the broader sense: a classic example of private judgment evaluating educational institutions.

Flcxner's approach differed from most contemporary accreditation efforts in two other significant ways. First, Flexner was not a member of the profession whose efforts he presumed to judge. An educator with no pretense of medical expertise, Flcxner nonetheless ventured to judge the quality of medical training in two nations. He argued that common sense was perhaps the most relevant form of expertise:

> Time and time again it has been shown that an unfettered lay mind, is. . . best suited to undertake a general survey. . . . The expert has his place. to be sure; but if I were asked to suggest the most promising way to study legal education, I should seek a layman, not a professor of law; or for the sound way to investigate teacher training, the last person I should think of employing would be a professor of education. (Flexner, 1960, p. 71)

It should be noted that Flexncr's point was only partially supported by his own study. Although he was a layman in terms of medicine, he was an *educator,* and his judgments were directed at medical *education,* not the practice of medicine, so even here appropriate expertise seemed to be applied.

Second, Flexner made no attempt to claim empirical support for the criteria or process he employed because he insisted that the standards he used were the "obvious" indicators of school quality and needed no such support. His methods of collecting information and reaching judgments were simple and straightforward: "A stroll through the laboratories disclosed the presence or. absence of apparatus, museum specimens, library, and students; and a whiff told the inside story regarding the manner in which anatomy was cultivated" (Flexner, 1960, p. 79).

Third, Flexner dispensed with the professional niceties and courteous criticisms that seem to typify even the negative findings yielded by today's accreditation processes. Excerpts of his report of one school included scathing indictments such as "Its so—called equipment is dirty and disorderly beyond description. Its outfit in anatomy consists of a small box of bones and the dried-up, filthy fragments of a single cadaver. A cold and rusty incubator, a single microscope,. ..and no access to the County Hospital. The school is a disgrace to the state whose laws permit its existence" (Flexner, 1910, p. 190).

Although an excellent example of expertise—oriented evaluation (if expertise as an educator, not a physician, is the touchstone), Flexner's approach is much more like that of contemporary educational evaluators who see judgment as the *sine qua liOn* of evaluation and who see many of the criteria as obvious extensions of common sense (for example, Scriven, 1973). But today's educational accreditation systems seem for the most part to have grown up differently. Whereas Flexner's review used the same process and standards for all medical schools reviewed, there is much more variability in contemporary national and regional accreditation systems in education. Agencies in the United States, such as the North Central Association (NCA) for accrediting secondary schools or the National Council for the Accreditation of Teacher Education (NCATE), have developed dualistic systems that include some minimum standards deemed important for all schools, along with an internal self—study component in which institutions can present their unique mission and goals, defend their reasonableness and importance, and report on how well that self-study approach is accomplishing its goals and what capabilities **it** offers for the foreseeable future. These two facets of accreditation are emphasized to greatly different degrees in various accreditation systems, leading Kirkwood (1982) to criticize accreditation for lacking "similarity of aims, uniformity of process, or comparability among institutions" (Kirkwood, 1982, p. 9), whereas others complain that the imposition of external standards by accrediting agencies denies institutions the opportunity of developing unique strengths.

Current accreditation systems also depend on the assumption that only members of a profession are qualified to judge the activities of their peers. Not only are

• accreditation site—visit team members drawn from the profession or occupation whose work they will judge, but also the standards and criteria are developed solely by members of that professional fraternity.32 For example, "The standards of techniques for accreditation of schools of teacher education have been determined by committees, comprised mainly of practicing teachers and teacher educators" (Floden, 1983, p. 262).

Although accreditation has historically focused on the adequacy of facilities,

qualifications of faculty, and perceived appropriateness of the educational design and processes used, rather than to assess the educational status of school graduates, several current accreditation systems aspire to justify their criteria and standards on the basis of empirical links to performance of program graduates (for example, Dickey & Miller, 1972; Study Commission, 1976). In large part such efforts are reactions to critics of accreditation who have typified accreditation as (in private correspondence from an unnamed colleague) "a bunch of anachronistic old fogies who bumble about with meters, measuring lighting and BTUs and counting the ratio of children per toilet, but failing to measure anything which could be conceived by any stretch of the imagination as related to what children are learning." Though obviously far overdrawn, such a caricature strikes a sensitive nerve among people responsible for accreditation systems, and such criticisms may account, at least in part, for a gradual de—emphasis on such quantitative indicators as square footage per student or number of volumes in the library and a move toward more qualitative indices dealing with purposes of schooling.

As accreditation systems have matured, they have taken on commonalities that extend to the accreditation of most primary, secondary, and professional schools, permitting Scriven (1984) to describe the distinctive features of contemporary accreditation as including (1) published standards; (2) a self—study by the institution; (3) a team of external assessors; (4) a site visit; (5) a site—team report on the institution, usually including recommendations; (6) a review of the report by some distinguished panel; and (7) a final report and accreditation decision by the accrediting body. Although not every accrediting system follows this prescription completely,33 this is an excellent description of most accreditation systems operating in education today.

Although viewed by some (for example, Scriven, 1984) as not truly an evaluative system, others (see Orlans, 1971, 1975) see accreditation as very much evaluative. Regardless of which view one holds, most would agree that accreditation has played an important role in educational change. Publication and application of the *E:'aluatiz'e Critcria* (National Study of School Evaluation. 1978) has helped shape recent changes in secondary schools in the United States. It is true that accrediting agencies have little real power over agencies who fail to take their recommendations seriously, as Floden states:

lfaccreditors instruct an institution to make particular changes, three options arc Open. First, officials may amass the necessary funds and make the changes. Second. they may decide the changes cannot be made and close their doors. Third. they may decide not to worry about what the accreditors say and make no changes. If an institution exercises

either of the first two options, the aims of accreditation have been realized. When the third option is taken, the process of accreditation has failed to achieve its main purpose. (Floden, 1983, p. 268)

Yet in our experience, the third option is only infrequently exercised. Fully accredited status is, if nothing more, a symbol of achievement highly valued by most educational institutions. Although there may be much room for improvement in the accreditation process, it appears to be a formal review process that will

be with us for a long time and, ***if*** the upsurge of thoughtful analyses of accreditation issues, problems, and potential (for example, Kells & Robertson, 1980) is any indication, there is reason to be optimistic that its impact can be pc5sitivc.

Other Formal Review Systems. Despite the wide reach of accreditaticn, there are those who feel **it** is an incestuous system that often fails to police itself adequately. As House (1980) has stated:

Public disenchantment with professionally controlled evaluation is re&cted in *the* declining credibility of accrcditing agencies. At one time **it** *was* sufficient for an institution to be accredited by the proper agency for the public to be assured of its quality —but no longer. Parents arc not always convinced that the school program is of high quality when **it** is accredited by the North Central Association. ln addition, political control of accrediting acdvitics is shifting to state governments. (House, 1980, p. 238)

Because of such concerns over accreditation's credibility, coupled with the pervasive feeling that decisions "closer to home" are preferable, many state boards or departments of education arc conducting their own reviews of elementary, secondary, and professional schools. Although these typically supplement rather than supplant reviews by private accrediting bodies, and generally use similar review strategies, they seem of greater consequence to institutions and programs reviewed because negative reviews result not only in loss of status but also possible loss of funding or even termination.

Consider, for example, the Utah State Board of Regent's system for reviewing academic departments within the state's tax—supported universities and colleges.34 The process and the structure for conducting these reviews have been formalized, general standards exist, a review schedule (every seven years) has been established, teams of experts in the academic discipline (also some "outsiders" to avoid bias and promote breadth of perspective) are used, and the results can influence both the future funding for the department and even its continued existence (though that is a rare outcome, indeed).

Newly proposed personnel evaluation procedures in education, such as competency tests for teachers, could eventually become institutionalized evaluative systems of a sort should they survive the political crossfire that makes any teacher-evaluation *system,* an easy target for a demanding but uninformed public.

Informal Professional Review Systems

Many professional review systems have a structure and set of procedural guidelines and use multiple reviewers. Yet some lack the formal review system's published standards or specified review schedule. For example, in the United States, state departments of education were required by federal law, over a period of many years, to establish a system to evaluate all programs and projects funded under a specific funding authorization designed to increase innovation in schools. Compliance varied widely, but those states that conscientiously complied established an evaluation structure in which districts receiving such funds were reviewed by teams of experts on a scheduled basis (for example, annually), with the results used to determine future funding levels and program continuation.35

Other examples of informal expertise—oriented review systems include review of professors for determining rank advancement or tenure status. Such reviews generally follow institutionalized structures and procedures, include input from several professional peers, and certainly can influence the status of the individual professor. Sometimes review schedules are set (for instance, some university policies require annual reviews of individuals in any rank lower than professor for five years or more), but usually the timing is dependent on the applicant's petitioning for review whenever he or she feels prepared. Reviewers rarely use pre— specified, published standards or criteria. Such standards are developed by each rank and promotion committee (a group of expert peers) within broad guidelines specified by the university and department and possibly by an existing statement of the role the individual is expected to play in contributing to those goals.

A graduate student's supervisory committee, composed of experts in the student's chosen field, is another example of an informal system for conducting expert—oriented evaluation. Structures exist for regulating such professional reviews of competence, but the committee members determine the standards for judging each student's preparation and competence. Few would question whether results of this review system affect the status arid welfare of graduate students.

Some may consider the systems for obtaining peer reviews of manuscripts submitted to professional periodicals to be examples of informal professional review systems. Perhaps. Many journals do use multiple reviewers, chosen for their expertise in the content of the manuscript, and sometimes empaneled to provide continuity to the review board. In our experience, however, the review structure and standards of most professional organs shift with each appointment of a new editor, and reviews occur whenever manuscripts are submitted, rather than on any regular schedule. In some ways, journal reviews may be a better example of the ad hoc professional review process discussed below.

Ad Hoc Panel Reviews

Unlike the ongoing formal and informal review systems discussed above, many professional reviews by expert panels occur only at irregular intervals, when circumstances demand. Generally, these reviews are related to no institutionalized structure for evaluation and use no predetermined standards. Such professional reviews are usually "one-shot" evaluations prompted by a particular. time—bound need for evaluative information. Of course, a particular agency may. over time, commission many ad hoc panel reviews to perform similar functions without their collectively being viewed as an institutionalized review system.

Funding Agency Review Panels. Many funding agencies use peer-review panels to review competitive proposals. Reviewers read and comment on each proposal and meet as a group to discuss and resolve any differences in their various perceptions.37 Worthen (1982) provided a manual of proposal review guidelines and instruments for use by external review panels, including (1) preparing for the proposal review (selecting reviewers, structuring review panels, preparing and training reviewers in the use of review instruments); (2) conducting he proposal

review (individual evaluation procedures, total panel evaluation procedures, methods for eliminating bias); and (3) presenting results of proposal reviews (summarizing review results). Justiz and Moorman (1985) and Shulman (1985) have also discussed particular proposal review procedures that depend upon the professional judgment of panels of experts.

Blue-Ribbon Panels. A prestigious "blue-ribbon panel," such as the National Commission on Excellence in Education, which was discussed in Chapter 1, is an example of an ad hoc review panel. Members of such panels are appointed because of their experience and expertise in the field being studied. Such panels are typically charged with reviewing a particular situation, documenting their observations, and making recbmmendations for action. Given the visibility of such panels, the acknowledged expertise of panel members is important if the panel's findings are to be credible. On more local scales, where ad hoc review panels are frequently used as an evaluative strategy on almost all types of educational endeavors, expertise of panel members is no less an issue, even though the reviewers may be of local or regional repute rather than national renown.

Although recommendations of ad hoc panels of experts may have major impact, they also may be ignored, for there is often no formalized body charged with the mandate of following up on their advice.

Ad Hoc Individual Reviews

Another form of expertise—oriented evaluation in education resides in the ubiquitous individual professional review of any educational entity by any individual selected for her expertise to judge its value. Employment of a consultant to perform an individual review of some educational program or. activity is commonplace. Such expert review is a particularly important process for evaluating educational textbooks, instructional media products, educational tests, and the like. Such instructional materials need not be reviewed on site, but can be sent to the expert, as suggested by Welch and Walberg (1968). A good example is the review of commercially available tests used by the Buros Institute (see Mitchell, 1985).

Educational Connoisseurship and Criticism

The roles of the theater critic, art critic, and literary critic are well—known and, in the eyes of many, useful roles. Critics are not without their faults (as we shall discuss later), but they are good examples of direct and efficient application of expertise to that which is judged. Indeed, few evaluative approaches in education are likely to produce such parsimonious and pithy portrayals as that of one Broadway critic who evaluated one new play with a single-line summary:

"The only thing wrong with this play is that it was performed with the curtain up!"

Although not championing one-line indictments, Eisner (1975, 1976, 1979a, 1979b) does propose that educators, like critics of the arts, bring their expertise to

bear in evaluating the quality of education. Eisner does not propose a scientific paradigm but rather an artistic one, which he sees as an important qualitative, humanistic, "nonscientific" supplement to more traditional inquiry methods.

Eisner (1975) has written that this approach requires *educational connoisseurship* and *educational criticism.* Connoisseurship is the art of appreciation—not necessarily a liking or preference for that which is observed, but rather an awareness of its qualities and the relationships among them. The educational connoisseur, in Eisner's view, is aware of the complexities in educational settings and possesses refined perceptual capabilities that makes the appreciation of such complexity possible. The connoisseur's perceptual acuity results largely from a knowledge of what to look for (advance organizers, or critical guideposts), gained through a backlog of previous relevant experience.

The analogy of wine tasting is used by Eisner (1975) to show how one must have a great deal of experience39 to be able to distinguish what is significant about a wine, using a set of techniques to discern qualities such as body, color, bite, bouquet, flavor, aftertaste, and the like, to judge its overall quality. The connoisseur's refined palate and "gustatory memory" of other wines tasted is what enables him or her to distinguish subtle qualities lost on an ordinary drinker of wine and to render judgments, rather than mere preferences. Connoisseurship does not, however, require a public description or judgment of that which is perceived, for the latter moves one into the area of criticism.

"Criticism is the art of disclosing the qualities of events or objects that connois—seurship perceives" (Eisner, 1979a, p. 197) as when the wine connoisseur either returns the wine or leans back with satisfaction to declare it of acceptable, or better, quality. Educational evaluators are cast as educational critics whose connoisseurship enables them to give a public rendering of the quality and significance of that which is evaluated. Criticism is not a negative appraisal, as Eisner presents it, but rather an educational process intended to enable individuals to recognize qualities and characteristics that might otherwise have been unnoticed and unappreciated. Criticism, to be complete, requires description, interpretation. and evaluation of that which is observed. "Critics are people who talk in special ways about what they encounter. In educational settings criticism is the public side of connoisseurship" (Eisner, 1975, p. 13). Educational evaluation, then, becomes educational criticism. The evaluator is the "instrument," and the data-collecting, analyzing, and judging are largely hidden within the evaluator's mind, analogous to the evaluative processes of art criticism or wine tasting. As a consequence, the expertise—craining, experience, and credentials—of the evaluator is crucial, for the validity of the evaluation depends on her perception. Yet, different judgments from different critics are tolerable, and even desirable, for the purpose of criticism is to expand perceptions, not to consolidate all judgments into a single definitive statement. -

Kelly (1978) has also likened educational evaluation to criticism by using literary criticism as his analogy. Although different in some features from Eisner's approach, it is similar enough to be considered as another example of this expertise—oriented evaluation approach.

HOW THE EXPERTISE-ORIENTED EVALUATION APPROACH HAS BEEN USED

As we noted earlier, this evaluation approach has been broadly used by both national and regional accreditation agencies. Two rather different types of accreditation exist. One is *institutional accreditation,* where the entire institution is accredited, including all of its more specific entities and activities, however complex. In essence, such institutional endorsement means the accrediting body has concluded that the educational institution, in general, meets acceptable standards of quality. The second type is *specialized or program accreditation,* which deals

• with various subunits in an institution,4 such as particular academic or professional training programs. As Kirkwood has noted, " institutional accreditation is not equivalent to the specialized accreditation of each ofhe several programs in an

• institution" (Kirkwood, 1982, p. 9). Rather, specialized accrecliting processes are usually more specific, rigorous, and prescriptive than are those used in institutional accreditation. Most specialized accreditation bodies are national in scope and frequently are the major multipurpose professional associations (for example, the American Psychological Association or the American Medical Association), whereas institutional accreditation is more often regional and conducted by agencies that exist solely or primarily for that purpose (for example, in the United States, the North Central Association or the New England Association).

A good example of how accreditation by private professional agencies and government-sponsored professional reviews are combined comes from Bernhardt's (1984) description of the evaluation processes of state, regional, and national education agencies, which collectively oversee reacher education programs in California:

Colleges and Universities in California must be accredited or approved by at least three agencies to offer approved programs of teacher education. Private institutions must first have the approval of the State Department of Education's Office of Private Postsecondary Education (OPPE) to offer ***degree*** programs. Public institutions must be authorized by their respective California State University and University of California systems. Second, institutions must be accredited by the Western Association of Schools and Colleges (WASC). Then, institutions must submit a document that states that the program is in compliance with all CTC guidelines in order to gain the approval of the Commission on Teacher Credentialing (CTC).

In addition to OPPE, WASC, and CTC accreditation, educational institutions often choose to be accredited by the National Council for Accreditation of Teacher Education (NCATE). (Bernhardt, 1984. p. 1)

Yes, formal professional review systems are alive and well, at least in California. Other uses of expertise—oriented evaluation are discussed by House (1980), who notes the upsurge of university internal—review systems of colleges, departments, and programs. He notes that such professional reviews are not only useful in making internal decisions and reallocating funds in periods of financial austerity but also may deflect suggestions that such programs should be reviewed by higher-education boards.

Uses (and abuses) of peer review by governmental agencies have been discussed by scholars in many disciplines (see Anderson, 1983; Blanpied & Borg, 1979; Gustafson, 1975). Justiz and Moorman (1985) listed several suggestions one panel of experts made for how the United States National Institute of Education might improve its peer-review processes: (1) identify individuals with recognized expertise and impeccable. credentials to serve as members of review panels; (2) provide for continuity and stability of review procedures and review groups by staggering memberships on continuing review committees; and (3) separate the scientific and technical review of proposals from general educational and social goals (adopting a two-tier system similar to the successful system used by the United States National Institute of Health). That some of these recommendations seem obvious is itself a commentary on the **need** f6r more thoughtful attention on the part of governmental agencies to the use of expert review as an evaluation approach.

Some funding agencies have also used panels of prestigious educators to evaluate the agencies to which research and development awards had been made. For example, the United States Office of Education empaneled review teams to visit and evaluate each member within its federally funded network of regional laboratories and university-based research and development centers, even though the evaluation focused on only some important outcomes (see Stake, 1970).

As for uses of Eisner's educational criticism approach, we are familiar with few applications beyond those studies conducted by his students (Alexander. 1977; McCutcheon, 1978; Vallance, 1978).

STRENGTHS AND LIMITATIONS OF THE EXPERTISE-ORIENTED EVALUATION APPROACH

Collectively, expertise—oriented approaches **to** evaluation have emphasized the central role of expert judgment and human wisdom in the evaluative process and have focused attention on such important issues as whose standards (and what degree of publicness) should be used in rendering judgments about educational programs. Conversely, critics of this approach suggest that it often permits evaluators to make judgments that reflect little more than personal biases. Others have noted that the *presunied* expertise of the reviewers is a potential weakness.

Beyond these general observations, the various types of expertise-oriented evaluation approaches have their own unique strengths and weaknesses. Formal review systems such as accreditation have several perceived advantages. Kirkwood (1982) lists accreditation's achievements:

(1) in fostering excellence in education through development of criteria and guidelines for assessing institutional effectiveness; (2) in encouraging institutional improvement through continual self-study and evaluation; (3) in assuring the academic community, the general public, the professions, and other agencies that an institution or program has clearly

defined and appropriate educational objectives, has established conditions to facilitate their achievement, appears in fact to be achieving theni substantially, and is so organized, staffed, and supported that it can be expecud to continue doing so; (4) in providing

counsel and assistance to established and developing institutions; and (5) in protecting institutions from encroachments that might jeopardize their educational effectiveness or academic freedom. (Kirkwood, 1982, p. 12)

The thoroughness of accreditation agencies has prevented the sort of oversimplification that can reduce complex educational phenomena to unidimensional studies. Other desirable features claimed for accreditation include the external perspective provided by the use of outside reviewers and relatively modest cost.

Of all these advantages, perhaps the most underrated is the self-study phase of most accreditation processes. Although it is sometimes misused as a public relations ploy, self-sudy offers potentially great payoffs, frequently yielding far more important discoveries and benefits than does the later accreditation site visit. Together. internal self—study and external review provide some of the advantages of an evaluative system that includes both formative and summative evaluation.

Formalized review systems also have nontrivial drawbacks. We have already commented on public concerns over credibility and increasing public cynicism that professionals may not police their own operations very vigorously. Scriven has called accreditation "an excellent example of what one might with only slight cynicism call a pseudo—evaluative process, set up to give the appearance of self— regulation without having to suffer the inconvenience" (Scriven, 1984, p. 73). The steady proliferation of specialized accrediting agencies suggests that there may indeed be truth to the suspicion that such processes are protectionist, placing professional self—interest before that of the educational institutions or publics they serve. Further, proliferation of review bodies, whether for reasons of professional self—interest or governmental distrust of private accreditation processes, can place unbearable financial burdens on educational institutions. Bernhardt (1984) suggests that the California system, which was described earlier, is too expensive to operate under current budgets, that it is not efficient, and that it is effective only for determining institutional compliance—not educational quality. Perhaps one accreditation visit may be relatively cost—efficient, as noted above, but multiple reviews can boost costs to unacceptable levels.41 This concern is exacerbated if one credits findings (see Guba & Clark, 1976) that seem to suggest that contemporary accreditation procedures have not demonstrated much effectiveness in changing or eliminating poor—quality institutions.

Scriven (1984) has cited several problems with accreditation: (1) no suggested weightings of a "mishmash" of standards ranging from trivial to important, (2) fixation on goals that may exclude searching for side effects, (3) managerial bias that influences the composition of review teams, and (4) processes that preclude input from the institution's most severe critics.

Informal peer-review systems and ad hoc professional reviews reflect many of the advantages and disadvantages discussed above for accreditation. In addition, they possess unique strengths and limitations. Some pundits have suggested that such expert reviews are usually little more than a few folks entering the school without much information, strolling through the school with hands in pockets, arid leaving the school with precious little more information, but with firm

conclusions based on their own preconceived biases. Such views are accurate only for *misuses* of expert-oriented evaluations. Worthen (1983) and Worthen and White (1986) have shown, for example, how ad hoc panel on—site reviews can be designed to yield the advantages of cross-validation by multiple observers and interviewers, while still maximizing the time of individual team members to collect and summarize a substantial body of evaluative information in a short time span Such ad hoc review panels can also be selected to blend expertise in evaluation techniques with knowledge of the program, and to avoid the naive errors that occur when there is no professional evaluator on the review team (Scriven, 1984).

Disadvantages of expert—oriented peer reviews include the public suspicion that review by one's peers is inherently conservative, potentially incestuous, and subject to possible conflict of interest. If evaluators are drawn from the ranks of the discipline or profession to be evaluated, there -are decided risks. Socialization within any group tends to blunt the important characteristic of detachment. Assumptions and practices that would be questioned by an outsider may be taken for granted. These and other disadvantages led us (Worthen, 1974b; Worthen & Sanders, 1984) to point to serious problems that can occur if a program is evaluated only by those with expertise in program content.

House has noted that confidentiality can be another problem because professionals are often loathe to expose their views boldly in the necessary public report. This normally results, he says, in "two reports, one an inside confidential report revealing warts and blemishes, the 'real' report, and a public report which has been edited somewhat. This dual reporting seems to be necessary for professional cooperation, but of course it makes the public distrustful" (House. 1980. pp.
240—241).

Obviously, the question of interjudge and interpanel reliability is relevant when using expert-oriented evaluations because so much depends on the professionalism and perception of the individual expert, whether working alone or as a team member. The question of whether a different expert or panel would have made the same judgments and recommendations is a troublesome one for advocates of this approach, for by its very definition, replicability is not a feature of expertise— oriented studies. Moreover, the easy penetration of extraneous bias into expert judgments is a pervasive concern.

Finally, the connoisseurship—criticism approach to educational evaluation shares, generally, the strengths and limitations of the other expertise—oriented evaluation approaches summarized above, in addition to possessing unique strengths and weaknesses. Perhaps its greatest strength lies in translating educated observations into statements about educational quality. Prior training, experience, and "refined perceptual capabilities" play a crucial role in every expertise—oriented approach to evaluation, but they are perhaps best explicated in Eisner's connoisseurship—criticism approach. One-cannot study his proposals and still lampoon expertiseoriented evaluation as a mere "hands—in—pocket" stroll through the school.

The connoisseurship—criticism approach also has its critics. House (1980) has cautioned that the analogy of art criticism is not applicable to at least one aspect of educational evaluation:

ft is nut unusual for an art critic to advance controversial views—the reader can choose to ignore them. In fact, the reader can choose to read only critics with whom she agrees. A public evaluation of an educational program cannot be so easily dismissed, however.

Some justification—whether of the critic, the critic's principles, or the criticism—is necessary. The demands for fairness and justice are more rigorous in the evaluation of public programs. (House. 1980. p. 237)

R. Smith (1984) *is* perhaps the harshest critic of the "educational crificism" approach to evaluation, fearing that "educational criticism will be esteemed more for its quality as literature and as a record of personal response than for its correct

• estimates of educatiqnal value" (R. Smith, 1984, p. 1). He continues by attacking

• Eisner's conception of educational criticism on philosophical and methodological grounds, and two of Smith's points are germane here. First, he quarrels with

• Eisner's contention that educational connoisseurs require no special preparation for their role by noting that anyone wishing to be an educational connoisseur—critic must possess the skills of literary criticism, knowledge of the theories of the social sciences, and knowledge of the history and philosophy of education, as well as sensitivity and perceptiveness—no small feat for the person whose primary training may be, for example, in the teaching of mathematics. How many could really qualify as educational connoisseurs is an important question. Second, Smith questions whether the same methodology is useful for judging the wide array of objects Eisner includes as potential objects of criticism. "Do the same nondis— cursive techniques serve the criticism of classroom life, textbooks, and school furniture?" (R. Smith, 1984, p. 14). Probably not.

APPLICATION EXERCISE

The Metropolitan Community Action Organization of Los Angeles received federal funds to establish a one-year education program for adults who have been unable to find employment for 18 consecutive months. A program was implemented that had two major components: (1) the teaching of basic skills such as reading, mathematics, and English as a foreign language, and (2) the teaching of specific vocational skills such as typing, shorthand, key punching, and drafting. The program was designed by adult-education specialists from a local university and representatives of the educational task forces of local unions.

Adults were tested as they entered the program by using standardized batteries in reading and mathematics. Entrants scoring below a grade equivalent of 8,0 were assigned to appropriate levels of reading and/or mathematics instruction, individual instruction was also provided for students who were not comfortable using the English language. Vocational offerings varied and depended on the unions' assessment of potential job openings in the Los Angeles area. Many of the vocational classes were held in the premises of places of business or industry. A few were conducted in the facility provided for the adult education program.

Using what you have learned about expertise-oriented evaluation approaches, indicate how these approaches might be used in the evaluation of this program.

What purposes could they serve? What could they contribute that other approaches might neglect or not address well? What process and criteria would you use to select your experts and to evaluate their performance?

SUGGESTED READINGS

EISNER, E. W. (1979a). *The educational imagination: On the design and evaluation of school programs.* New York: Macmillan.

FLODEN, R. E. (1980). Flexner, accreditation, and evaluation. *Educational Evaluation and Policy Analysis,* 20, 35—46.

KELLS, H. R., & **ROBERTSON,** M. P. (1980). Post-secondary accieditation: A current bibliography. *North Central Association Quarterly,* 54, 411—426.

KIRKWOOD, R. (1982). Accreditation. In H. E. **MITZEL** (Ed.), *Encyclopedia of educational research* (Vol. 1, 5th Ed.). New York: Macmillan and The Free Press.

National Study of School Evaluation (1978). *Evaluative criteria.* Arlington. VA: Author.

intendancy? Who will know that the harsh evaluation of a computer-assisted math curriculum stemmed more from the evaluator's aversion to computers than from any of the curriculum's attributes? In short, the notion that any evaluator can be a paragon of impartiality is naive. The best that any evaluation approach can hope for is to control bias sufficiently so that **it** does not significantly distort or alter results.

Where most evaluation approaches attempt to reduce bias, the *adversary-oriented approach* aspires to balance it, attempting to assure fairness by incorporating both positive and negative views into the evaluation itself. We would consider an evaluation adversary—oriented if both sides of issues or questions were argued, one side by advocates (those in favor) and the other by adversaries (those opposed). Various types of data (ranging from test scores to human testimony) might be selected and used by either side as evidence to support its arguments. Generally some type of hearing would be held so that the opposing views could be presented and debated before whoever would serve as "judge" or "jury" to decide on the relative merits of the opposing cases. There would be no presumption that the proponents and opponents of a curriculum being evaluated would be unbiased in appraising **it.** On the contrary, we would expect their biases to surface as they mounted their respective defenses of or attacks on, the curriculum. And by encouraging biases on both sides to surface, we help ensure a balanced method of gathering information regarding the curriculum.

Adversary-oriented evaluation, then, is a rubric encompassing a collection of divergent evaluation practices that might loosely be referred to as *adi'ersarial* in nature. In its broad sense, the term refers to all evaluations in which there is planned opposition in the points of view of different evaluators or evaluation teams—a *planned* effort to generate *opposing* points of view *u'ithin* the overall evaluation.42 One evaluator (or team) serves as the program's advocate, presenting the most positive view of the program possible from the data, while another evaluator (or team) plays an adversarial role, highlighting any extant deficiencies in the program. Incorporation of these opposing views within a single evaluation reflects a conscious effort to assure fairness and balance and illuminate both strengths and weaknesses of the program. As Levine (1982) has put it,

In essence, the adversarial model operates with the assumption that truth emerges from a hard, but fair fight, in which opposing sides, after agreeing upon the issues in contention.

present evidence in support of each side. The fight is refrced by a neutral figure, and all the relevant evidence is weighed by a neutral person or body to arrive at a fair result.

(Levine, **1982,** p. 270)

Several types of adversarial proceedings have been invoked as models for adversary evaluations in education, including judicial proceedings, congressional and other hearings, and structured debates, each of which we shall consider in this chapter.

DEVELOPERS OF ADVERSARY-ORIENTED EVALUATION APPROACHES AND THEIR CONTRIBUTIONS

Adversary-oriented evaluation approaches can subsume, draw from, and be incorporated within other evaluation approaches. For example, there is considerable dependence on expert-oriented evaluation (discussed in Chapter 8) in many adversary proceedings (for example, the use of expert witnesses in trials and congressional hearings). Adversary evaluation also shares with the evaluation approaches discussed in Chapter 10 dependence on multiple perspectives about what is evaluated. (lnaeed, one such approach, transactional evaluation, proposes the use of proponents and opponents of planned changes on evaluation teams charged to study innovations.) We distinguish adversary evaluation, however, b its use of planned, structured opposition as the primary core of the evaluation and by its derivation from metaphors drawn from more venerable adversarial paradigms.

Origins of Adversary-Oriented Evaluation. Rice (1915) proposed an evaluation method intended to eliminate graft and increase governmental efficiency by presenting facts about waste and corruption to a mock "judge and jury." Although only partly aimed at education, Rice's approach is the first proposed use of "adversary evaluation" with which we are familiar. The idea was not further developed, however, for 50 years. Guba (1965) suggested that educational evaluation might use aspects of the legal paradigm. If trials and hearings were useful in judging truth of claims concerning patents and products, and if human testimony were judged acceptable for determining life or death, as in the judicial syster, then might not legal proceedings be a useful metaphor for educational evaluation? Might there be merit in evaluation "trials," in taking and cross—examining human testimony, and in using the concept of advocacy to assure that evaluation fairly examined both sides of issues?

At first, Guba's ideas seemed to fall on deaf ears, for that was the era when evaluators were about the business of refining the application of social science research methods (for example, experimental design) to educational evaluation, as well as developing promising new approaches drawn from other relevant paradigms (management-oriented approaches based on decision theory and systems analysis). But gradually a few colleagues began to test the utility of Guba's suggestions.

The first self-conscious effort to follow a particular adversary paradigm was made in 1970 by Owens. Designed to test the usefulness of a modified judicial model, the evaluation focused on a hypothetical curriculum and included pretrial conferences, cases presented by the "defense" and "prosecution," a hearings officer, a "jury" panel of educators, charges and rebuttals, direct questioning and redirected questions, and summaries by the prosecution and defense. The reports (Owens. 1971, 1973) were intriguing to the community of educational evaluators and led to further conceptual and empirical work on the adversary approach (for example. Kourilsky, 1973; Wolf, 1973, 1975; Levine, 1974; Stake & Gjerde, 1974;

Kourilsky & Baker, 1976; Owens & Hiscox, 1977; Worthen & Owens, 1978; Levine & Rosenberg, 1979; Owens & Owen, 1981; and House, Thurston. & Hand, 1984). Several of these studies involved what might best be termed *advocate- adversary evaluation,* where an advocate evaluator presents the most favorable review possible and an adversary evaluator presents the most critical and damaging case that might be made, but there are no adversarial interactions or rebuttals surrounding the two stated positions. °

As these efforts to develop the adversary approach continued, several evaluations occurred that could be judged truly adversarial in nature (for example, Hiscox & Owens, 1975; Wolf, 1975; Stenzel, 1975; Nafziger, Worthen, & Benson, 1977; Levine and others, 1978; Wolf, 1979; Brathwaite & Thompson, 1981; Madaus, 1981; Popham, 1981). These studies have used widely divergent styles, and reactions to them have been mixed (as will be discussed later).

In the balance of this chapter we shall consider three general approaches to adversary evaluation: (1) adaptations of the legal paradigm and other "two view" adversary hearings; (2) adaptations of quasi—legal and other adversary hearings where more than two opposing views are considered; and (3) use of debate and other forensic structures in adversary evaluations.

The Judicial Evaluation Model and Other 'Pro and Con' Adversary Hearings

The "fight theory" underlies most models of litigation for resolving differences among opposing parties. According to Auerbach, Garrison, Hurst. and Mermin, (1961), this theory holds that the facts in a case can best be determined if each side tries as hard as possible, in a keenly partisan spirit, to provide the court with evidence favorable to that side. Although not disagreeing with the advantages of this posture, Frank (1949) has cautioned that disadvantages occur when "the partisanship of the opposing lawyers blocks the uncovering of vital evidence or leads to a presentation of vital testimony in a way that distorts it" (Frank. 1949, p. 81). Efforts to adapt aspects of the legal paradigm for use in educational evaluation have attempted to capitalize on the potentials cited by Auerbach and colleagues while avoiding the pitfall of which Frank warns.

Owens (1973) listed several characteristics of the adversary proceeding that he believed made it more appropriate for educational evaluations than adaptations of more familiar models:

1. The rules established for handling the adversary proceedings arc quite flexible.
2. Complex rules of evidence are replaced by a free evaluation of evidence based solely upon whether the evidence is considered by the hearings officer to be relevant.
3. Both parties can be required before the trial to inform the hearings officer of all relevant facts, means of proof and names of witnesses.
4. A copy of the charges is furnished to the hearings officer and defendant

beFore the trial and the defendant has the option of admitting in advance to certain chargcs and challenging others.
5. Witnesses are allowed to testify more freely and to be cross—examined.
6. Experts are often called upon to testify even before the trial.
7. Pretrial conferences of the hearings officer with both parties tend to make the trial less a battle of wits and more of a search for relevant facts.
8. In addition to the two parties involved, other interested groups may be permitted to participate. (Owens, 1973, pp. 296—297)

Owens also indicated that adversary proceedings in education should not be used to replace existing designs for data collection and analysis, but rather to provide an alternative way of intefpreting, synthesizing and reporting evidence.

The work of Wolf *(1973*, 1975) has been particularly thoughtful in relation to how evaluators might better define evaluation issues, what role personal testimony might play in evaluation, procedures for direct questioning and cross—examination, and rules of admissibility of evidence. Borrowing concepts from both jury trials and administrative hearings, Wolf proposed the *Judicial evaluation mode!,* which included a statement of charges. opposing counselors, witnesses, a judge or hearings officer, and a jury panel. Four stages are proposed:

1. *Issue* ***generation:*** identification and development of possible issues to be addressed in the hearing
2. *Issue selection:* elimination of issues not at dispute and selection and further development of those issues to be argued in the hearing
3. *Preparation of arguments:* collection of evidence, synthesis of prior evaluation data to develop arguments for the two opposing cases to• be presented
4. *The hearing:* including prehearing discovery sessions to review cases and agree on hearing procedures, and the actual hearing's presentation of cases, evaluation of evidence and arguments, and panel decision.

Wolf (1975, 1979) made clear that his intention was merely to use the law as a metaphor for educational evaluation, not to replicate legal procedures. He was also prompted by critiques of problems m applying the legal paradigm to educational evaluation (for example, Popham & Carlson, 1977; Worthen & Rogers, 1977) to argue that his model was not an adversarial debate or adversary evaluation, as such: "the metaphors of law are just that—metaphors. ... Once the concepts are taken too literally, the object of judicial evaluation then becomes *winning*. This is precisely *not* what the JEM [judicial evaluation model] strives for" (Wolf, 1979, p. 22).

Levine and Rosenberg (1979) have provided an insightful examination of numerous issues in adapting legal analogs for use in evaluation (for example, the burden of proof and use of presumptive evidence). They point out that although adversary models such as jury trials, administrative hearings, appellate proceedings, and arbitration hearings all have unique ways of using evidence and argument, they also have important similarities, including (1) an existing controversy between two or more parties; (2) formal case presentation by advocates for each position; (3) facts heard and decision rendered by an impartial arbiter; and (4)

decision based solely upon argument heard and evidence presented during the proceeding.

Adversary Hearings with More than Two Opposing Views

Many committee hearings are not adversarial. Some of the review panels discussed in Chapter 8 (such as blue-ribbon panels) may hold public hearings to collect information pertinent to their charge. Appointed commissions charged with the resolution of controversial issues (for example, the National Commission on Excellence in Education described in Chapter 1) frequently hold hearings to obtain evidence and opinions relevant to their mission. House (1980) has cited as one such example the frequent use in England of commissions and councils headed by prominent citizens to provide guidance to government policymakers. Hearings held by most committees and commissions are decidedly not adversarial in structure, however, for no efforts re made to articulate or contrast opposing points of view.

Several other types of committee hearings are structured to identify and explore all the points of view represented in a particular context. Although not "adversarial" in the strict sense of the word, because, as Smith (1985) has noted, they explore a variety of positions, not just pro and con, we prefer to include them here because (1) they reflect multiple viewpoints, which often are in conflict with one another, thus perhaps qualifying as "multiadversarial" in nature: and (2) they frequently use hearing processes, questioning, cross—examination, interaction concerning alternate viewpoints, and summary statements of the various positions, all procedures typical of the two-sided "pro and con" adversary hearing. St. John (undated), in referring to such hearings as the "committee approach" to evaluation, listed as key characteristics the following:

• All of those with a stake in the evaluation—decision—makers, evaluators, program personnel, clients, and other interested persons—are brought together in the same place at the same time for a careful review of the issues at hand.

- A public hearing with testimony, questioning, cross-examination, and summary statements *produces a full exposition of evidence and illuniinates differiu, points of view about that evidence.*

• The committee hearing method consists of public, verbal, face-to-face interactions, and therefore generates a high degree of personal involvement. Consequently, committee hearings are likely to have a strong impact on those involved, as well as on those who observe them.

• Because interaction between different points of view takes place, a process of communication and education occurs, and the evaluation makes its impact as it is happening. (St. John, undated, p. 2; italics added)

St. John also suggested that committee hearings "... may be useful. . . when the impact of the evaluation and its follow—through dependS on the *consensus of multiple perspectives,* and such consensus is unlikely without significant interaction" (p. 3, italics added). Had he said "presentation of multiple perspectives," we would agree

fully, for there arc obviously instances where consensus among disparate views is not attained, yet issues are resolved through hearing and weighing the evidence supporting those alternate viewpoints enroute to a decision that may or may not be consensual. We see the focus of such "adversary" committee hearings as the presentation and examination of multiple perspectives that illuminate all legitimate views prior to final resolution of the issues.

The most frequently proposed model for this type of adversary evaluation is the congressional or legislative investigative hearing (Stenzel, 1982; Levine and others, 1978). With origins nearly as old as the origins of parliamentary process, congressional hearings seek to gain information or unveil truth. Although chief counsel and possibly minority counsel might be assigned to assist the committee, the viewpoints arc seldom dichotomous partisan views but rather reflect a broad spectrum of individual and group positions (witness the well—known example of the Watergate hearings). Ensuring that all these important views are heard **sometimes** requires special powers (for example, subpoenaing witnesses), which would seem more difficult to enact and possibly less appropriate in educational settings.

Adversary Debates and **Other Forensic Structures**

Several approaches that qualify as adversary-.oricnted do not employ hearing processes. For example, Kourilsky (1973) proposed that pro and con arguments be presented to a decision-maker, who would examine the evidence and question the presenters, ultimately arriving at the decision that seemed fair given, both positions. Kourilsky and Baker (1976) described an adversary model in which two teams prepared, respectively, affirmative and negative appraisals of that which was evaluated (the preparation stage), met to present the views to one another, cross—examining and critiquing one another's contentions on prespecified criteria (the confrontation stage), and engaged in open-ended discussions until reconciliation of views was attained and translated into written recommendations in a single report. Levine (1974) proposed that a resident adversary or critic might be assigned to a research project to challenge each bit of information collected, searching .for other plausible explanations. The Stake and Gjerde (1974) strategy of having two evaluators prepare separate reports summing up opposing positions for and against the program is yet another variant adversarial approach that does not depend on a hearing format. Donmoyer (undated) proposed a "deliberative" approach to evaluation, which focused on assessing and balancing alternative conceptions of reality and the differing value positions underlying these conceptions. "Because deliberative evaluation is primarily concerned with fostering understanding of alternative conceptions of reality," the evaluator's role is "to foster interaction and facilitate communication *among* representatives of various stakeholder [groups]. .

(Donmoyer, undated, pp. 9—10). Donmoyer saw different world views as the cause of underlying disputes, which could be resolved by open presentation of alternative views in some t.ype of educational forum:

Through the process of communication, those who disagree can, in principle, at least, expand their understanding of an issue by viewing that issue from their opponents' perspectives. (Donmoyer, undated, p. 11)

Nafziger and others (1977) described an adversary evaluation design employing a modified debate model for presenting data collected in a comprehensive evaluation to ensure that both sides of controversial issues were illuminated. This model *was* used in an adversary evaluation of a statewide team—teaching program in Hawaii.

HOW THE ADVERSARY-ORIENTED EVALUATION APPROACH HAS BEEN USED

The Hawaii evaluation conducted by Nafziger atid his colleagues is the only example we know that made any effort to follow the debate model, as opposed to other forensic models. The program evaluated was a controversial "team— teaching" program. Two evaluation teams were formed, and once they had agreed on the basic design for the evaluation, they were randomly assigned positions as the program's advocate or adversary. Each team drew from the common data provided for in the original design and, in addition. was free to collect supplemental data. The teams wrote and exchanged reports and then prepared written rebuttals. Finally, the team leaders presented their reports and arguments verbally and rebutted their opponents' arguments in a standard debate format, before influential Hawaiian educational, governmental, and private leaders. who were given opportunities to ask questions of both team leaders.

The written final reports and live debates sparked great interest, including wide viewing of two television airings of an hour-long condensed version of the debate. Further, Hawaiian decision—makers were very favorable toward the adversarial format of the evaluation (Wright & Sachse, 1977). Despite such receptivity, and selection of this evaluation by the American Educational Research Association as the best all—around evaluation study of 1977, some participants in this study later expressed serious misgivings about aspects of this particular adversary approach (Worthen & Rogers, 1980) or about adversary evaluation in general (Popham & Carison, 1977).

Several adversary—oriented evaluations have incorporated aspects of the legal paradigm. Wolf's judicial evaluation model has been used in (1) the evaluation of Indiana University's undergraduate teacher education programs (Wolf. 1975: Arnstein, 1975); (2) examination of the United States Bureau of Education for the Handicapped's implementation of a law mandating that all handicapped children have available a free and appropriate education (Wolf & Tymitz, 1977); (3) studying policy formulation in a local school district (Wolf, 1978), and (4) a formative evaluation of the effectiveness of "networking" among agencies in Virginia's employment and training program (Braithwaitc & Thompson, 1981). A modified version of the judicial evaluation model (which omitted the jury or panel whose purpose in previous applications was to make recommendations or decisions) was used in the highly publicized Clarification Hearings on Minimum Competency Testing sponsored by the United States National Institute of Edu—

cation (Hcrndon, 1980; National Institute of Education, 1981; Madaus, 1981; Popham, 1981).

Other adversary hearings employing legal methods include Owens (1971), described earlier in this chapter, and an evaluation of an experience—based career education program for high school students (Hiscox & Owens, 1975; Owens, Haenn, & Fehrenbacher, 1976), which produced a videotaped hearing presided over by a law professor serving as "judge," with professors of evaluation as the defense and prosecution "attorneys." In an evaluation of doctoral candidacy procedures in a university psychology program, a public hearing resembling a jury trial was used (Levine, 1976; Levine and others, 1978).

Aspects of both the'legal and debate models were employed in an evaluation of an experimental undergraduate program in liberal arts at the University of Illinois (Stenzel, 1976). A debate consisting of opening arguments, rebuttals, and final summaries of both advocate and adversary positions was presented to a panel of judges, following the appellate court hearing in which such a panel decides issues under contention.

Clyne (1982) summarized the uses of the adversary process in educational evaluation: (1) summative evaluation; (2) formative evaluation; (3) social science debate; (4) policy analysis and debate; (5) school governance and local decision making; and (6) issue resolution and policy formation. Worthen and Rogers (1980) reported that a survey of a group of key educators and policymakers showed most (81 percent) thought adversary evaluation was appropriate for .summative evaluation, whereas only 15 percent felt it should be used in formative evaluation. Braithwaite and Thompson (1981) disagreed, stating that their study showed adversary evaluation could also serve well in formative evaluation. In their evaluation of the national Clarification Hearings on Minimum Competency Testing, Estes and Demaline's (1982) surveys showed that most participants or potential users of such evaluations view adversarial approaches as more useful for summative than formative decisions.

STRENGTHS AND LIMITATIONS **OF** THE ADVERSARY-ORIENTED **EVALUATION APPROACH**

Some strengths and weaknesses transcend particular adversary approaches and speak to the merits of the adversarial concept itself. For example, most observers would agree that building opposing viewpoints into an evaluation tends to illuminate both the positive and negative aspects of an educational program better than most other evaluation approaches. Adversary approaches also tend to broaden the range of information collected. A strength common to all of the adversary approaches is the interest they create on the part of their intended audiences. Indeed, one of this approach's greatest strengths is that it can satisfy the audience's informational needs in an interesting, informative manner. Nearly everyone loves a contest.

Adversary—oriented evaluation is also sufficiently broad and pluralistic that it can

be combined with other approaches. For example, there is nothing to prevent the use of an expertise—oriented evaluation approach by both teams in an adversary—oriented study, any more than it would violate this approach for the advocate to use a participant-oriented approach while the adversary employed a management- oriented approach.

Openness to diverse viewpoints and open participation by stakeholders who might be excluded by most other approaches are other advantages. Further, this diversity increases the educative value of the hearings.

Another general advantage to adversary-oriented evaluation is that it anticipates (and largely blunts) the nearly inevitable criticisms offered by anyone whose position is not supported by the findings. It is difficult to argue that an evaluation is unfair if it examines and presents *both* sides of an issue. Use of adversary evaluation to diffuse political heat may be an unexpected fringe benefit of this approath.' Because opposing views are incorporated *into* the evaluation, most of the pros and cons are argued in an open forum, diverting much subsequent criticism of the evaluation itself. In short, there is more openness to examining issues rather than focusing on one particular point of view. This is consistent with findings from social psychology literature in persuasion and communication research (see Paulson, 1964) that suggest that opinions are modified more readily when both positive and negative views are reported.

Another advantage is the substantial, rigorous planning required of most adversary evaluations (no one wants to be humiliated by an opponent gloating over an easy victory). Few evaluation studies are so carefully planned as those with an adversary orientation.

Adversary evaluation also has, in a sense, a built—in "meta—evaluation": an evaluation of the evaluation. The collection, analysis, and interpretation of data used to support any point of view will be painstakingly criticized by those in opposition. All that remains is to do a more general meta—evaluation of the overall study.

The use of direct, holistic human testimony is frequently cited as a strength of adversary—oriented evaluations, as is cross-examination of that testimony.47 Considerable use can be made of expert witnesses, thus enabling experts to draw inferences that might elude "lay" educators. Testing of hidden biases is another strength of this approach, as is examination of alternative interpretations of evidence. Certain legal metaphors may be particularly useful. For example, the British judicial system's pretrial "exploration for discovery" provides an opportunity for opposing barristers to disclose to one another their cases and supporting evidence in the interest of finding any common ground. When two adversaries agree on any data, interpretation, or conclusion, it lends great credence to that aspect of the evaluation. The requirement that any evidence presented be understandable also precludes the type of obfuscation and educational "baffle-gab" that permeates many traditional evaluation reports.

To summarize the discussion thus far, we believe an adversary-oriented evaluation approach may be useful when (1) the object of the evaliation affects

many people, (2) controversy about the object of the evaluation has created wide interest, (3) decisions are summative, (4) evaluators are external, (5) clear issues are involved, (6) administrators understand the intensity of adversary-oriented evaluations, and (7) resources are available for additional expenses required by adversarial strategies.

Despite their potential for making evaluation findings more interesting and meaningful to educational decision-makers, adversary-oriented approaches to evaluation are not yet sufficiently well developed to serve as a standard or model for future efforts. As yet there is little beyond personal preference to determine whether such evaluations would best be patterned after jury trials, congressional hearings, debates, oç other adversarial arrangements. Preoccupation with the respective paraphernalia of these various approaches could cause evaluators to overlook the benefits that might accrue from use of adversary-oriented evaluation, namely, including planned opposition among evaluators. Despite its intriguing possibilities, we are not convinced that the legal paradigm is necessarily the best pattern.

Evaluators may forthrightly protest (for example, Wolf, 1975, 1979) that rigid adherence to a legal model is not intended, yet many continue clinging to the more trivial courtroom rituals that seem unnecessary or downright inappropriate in educational evaluation—what Owens (1973) has called "entanglement in legal technicalities." For instance, replicating the theatrical aspects of the courtroom in adversary hearings is distracting and has made a mockery of some educational evaluations. Cloaking the person presiding over an educational heating in a black robe seems as pretentious and inane as placing powdered wigs on senators presiding over congressional hearings.

Use of the legal paradigm can also result in a seductive slide into what might be termed an "indictment mentality," which can do a disservice both to evaluation efThrts and to the programs being evaluated. Adversary—oriented evaluation literature that invokes the legal model tends to use terms such as "statement of charges" (Hiscox & Owens, 1975), "defendant" (Levine & Rosenberg, 1979), "not guilty" (Levine, 1982), "trial by jury" (Wolf, 1975), and the like. That orientation may be appropriate when there is a formal complaint against an educational program, as in the occasional investigation of some education program for malfeasance, misuse of funds, or gross mistreatment of students. But such situations are rare; and formal complaints, plaintiffs, and litigants are conspicuously absent in the typical educational evaluation—and rightly so. Educational evaluation should aspire to improve educational programs, not determine their guilt or innocence. Although evaluators must of necessity render judgments of worth, judging merit is not the same thing as determining guilt.

It is not only the vocabulary of the legal model that is problematic but also its characteristic of serving only when there is a problem to be solved. There is already too much tendency to view evaluation as something done when a program is in trouble, when there is a crisis or failing that requires correction. It would be unfortunate if this tendency were exacerbated and evaluations conducted only when a complaint had been lodged, an accusation leveled, an offending program accused. It is precisely this orientation that we fear may be a side effect of basing

evaluations on the legal model, or on any model meant to be applied only in problem solving or crisis situations. It would be far more salutary if educators came to view evaluation as something routinely carried out to help them keep programs operating at maximum effectiveness and efficiency.

Of course, one can use aspects of the legal paradigm, such as cross—examination by adversaries, without requiring full or even partial courtroom procedures (as demonstrated by congressional hearings or interviews conducted jointly by partisan interviewers). Wolf (1975) and Hiscox and Owens (1975) have shown that one can adapt portions of the legal model without adopting it in its entirety.

Another general concern with adversary—oriented evaluation is whether it provides decision-makers with the full range of needed information. Presentation of strong pro or con positions might increase the probability of an extreme decision. In emphasizing polar positions, educators stand to both gain and lose. They may gain a broader spectrum of data and with more diverse interpretations provided to decision-makers; few other evaluation approaches seem likely to push as far in both positive and negative directions. But in broadening that spectrum, they may compromise the very neutrality so essential to rational decision making.

As mentioned earlier, decision—makers may place greater confidence in conclusions and recommendations agreed to by both sides. Although this seems patently sensible, experience with adversary evaluations suggests such agreement is unlikely to be a spontaneous byproduct of the sparring and jousting that often occurs between adversaries. Most adversary approaches have a competitive element; it is expected that one of the adversaries will win and the other lose. When competition is high, cooperation tends to be lower. Mutual agreements arc often abandoned in the adversaries' rush to turn every opposing argument to their own advantage. When winning is at stake, seemingly rational opponents question the obvious. And shared conclusions are not easily come by. Most adversary-oriented evaluation approaches could profit from a better mechanism for seeking and reporting areas of agreement.

Popham and Carlson (1977) point to "disparity in proponent prowess' as a deficit of adversary evaluation, claiming it is all too likely that the case will be decided because of a disparity in skill of the competing teams (or individuals), with the audience influenced more by the persuasiveness of the protagonists than- by the strength of the evidence that supports their arguments. The potential for a skilled orator without solid supportive data to sway the "jury" by eloquence alone is, unfortunately, a real possibility.

Assignment of adversaries to pro or con positions is also a difficult matter because of the possible biases that they might bring with them. As noted at the beginning of this chapter, one object of adversary—oriented evaluation is not elimination of bias but rather balancing and publicizing of that bias. Of course, biases are unlikely to be eradicated by assignment to a position. Imagine the plight of Ralph Nader if he were assigned to defend a program or product.

It may not be an explicit assumption, but many adversary—oriented evaluations proceed as if there were an unspoken *ob1iation* to present two *equcIly convincing*

cases, pro and con. Naturally, no one would tolerate an advocate who presented a weaker case than the data warranted; but what about one who erred in the other direction, feeling compelled to keep up with the opposition, even if it meant straining or ignoring the data? Such an orientation might be appropriate in a forensic society wherein the result of the debate seldom had much effect on the proposition being argued, but not in an evaluation where the outcome will influence real programs and real people.

Like the legal paradigm, the debate model also has irrelevancies that should be strained out before it is applied to education. The touchstones of debate are polemics and persuasion, not truth, which is central to the validity of evaluation studies. Debaters sur'ely use facts and cannot normally afford to ignore the evidence at hand. But seldom is the debater forced to adhere as tightly to the plain, unadorned facts as is the conscionable evaluator. Logic can provide a permissive climate for manipuiing data until the form is favorable. Probably more sophistry results from debats' perversions of syllogistic logic than any other form of self-deception. A skilled debater can often build a remarkably strong case on a very flimsy foundation.

Many commentators have pointed out that adversary-oriented evaluations are time-consuming and expensive, requiring extensive preparation and considerable investment of human and fiscal resources (for example, Owens, 1973; Stenzel, 1982; Popham & Carlson, 1977). Braithwaite and Thompson (1981) said the judicial evaluation model's most serious problem is that it is a "heroic model," requiring

> a large number of participants, many of whom are in important roles. We utilized four case presenters, seven panelists, .a hearing officer, a panel facilitator,, a nonparticipant
> observer, and two research assistants in addition to ourselves, 13 witnesses, and two speakers who provided contextual statements at the outset of the hearing. (Braithwaite & Thompson, 1981, p. 16)

Levine and others (1978) estimated that over 80 percent of their effort in an evaluation using an adaptation of a jury trial went into preparing the case and managing the process, and less than 20 percent went into the actual hearing. Kourilsky and Baker (1976) noted as a potential snare of adversary evaluation the temptation to report all of the voluminous information collected with this approach. In short, questions have been raised about whether the adversary approach to evaluation is worth its considerable costs.

The real question, however, is not cost, but cost—effectiveness or cost—benefit. On these dimensions it seems apparent that benefit must be argued on grounds that adversary evaluation increases representativeness of the data, fairness of
- the instruments, communication between evaluators and decision-makers, and identification of all the pros and cons. Whether adversary evaluation really provides additional benefits must remain an open question until someone sees fit to research the issue.

A final concern of critics of adversary—oriented evaluations is that those who serve as judges are fallible arbiters. Popharn and Carlson (1977) have worried about

the lack of a process for appealing unwise decisions by arbiters, stating that the lack of a "higher court of appeals" in educational evaluation precludes rectifying improper judgments.49 House (1980) echoes this concern and also lists as a criterion the contention that the "adversary model" may resolve conflicts but has limited potential for getting at the truth of a matter. He quotes Ramsey Clark, former United States Attorney General, as saying, "If there is a worse procedure for arriving at the truth, I don't know what it is."

APPLICATION EXERCISE

The curriculum council for a large school district decided that one of the major weaknesses of the elementary curriculum was its writing program. Junior high teachers reported that students were generally unable to write cohesive, descriptive paragraphs. On the recommendation of the council, six elementary schools were selected to participate in a pilot project to develop an elementary writing program. The nucleus of the developmental staff consisted of the faculty of these schools. This staff included also a specialist in creative writing from the local university and a representative from the curriculum council. The statT worked together for eight weeks in the summer to develop a program that would be used in all six grades. When school opened, they met twice weekly to discuss the way the course was progressing and to act on the recommendations **of** the evaluation team. The evaluation team had been appointed by the curriculum council and had the following responsibilities: (1) decide what questions should be asked of the program; (2) select the appropriate criteria for success of the program; and (3) gather information about the program and give it to the program staff with recommendations that they either improve, terminate, or maintain the program.

How could adversary-oriented evaluation approaches be used to address the evaluation needs of the council? Provide details about who would be involved, what the procedures would involve, what reports would be generated. and how the results might be used.

SUGGESTED READINGS

KOURILSKY, M. (1973). An adversary model for educational evaluation. *Ez'aluarion Comment,* 4(2), 3—6.

LEVINE, M. (1982). Adversary hearings, In N. L. **SMITH** (Ed.), *Co,nmuniation strate@es ill evaluation*. Beverly Hills, CA: Sage.

OWENS, T. R. (1973). Educational evaluation by adversary proceeding. In E. R. **HOUSE** (Ed.), *School evaluation: The politics and* ***process.*** Berkeley, CA: McCutchan. pp. 295—3D5.

STENZEL, N. (1982). Committee hearings as an evaluation format. In N. L. **SMITH** (Ed.), *Field assessments !f ilnol'atipe evaluation methods.* New Directions for Program Evaluation, No. 13. San Francisco: jossey—l3ass.

WOLF, R. L. (1975). Trial by jury: A new **evaluation** method. *Phi Delta Kappan. 57(3),* 185—187.

the educational community argued that the human element, which was reflected in the complexities of everyday reality and the different perspectives of those engaged in education, was missing from most educational evaluations.

Consequently, a new orientation to evaluation was born, one that stressed firsthand experience with educational activities and settings. This general approach, which grew quickly during the 1970s and 1980s, is aimed at observing and identifying all (or as many as possible) of the concerns, issues, and consequences integral to educational enterprise.

In large part a reaction to perceived deficits in other evaluation approaches, this orientation encompasses a wide variety of more specific proposals that might be generally tied together by their acceptance of the intuitionist-pluralist philosophy of evaluation (see Chapter 4). Most of those who contributed to the development and use of this evaluation approach exhibit a preference for naturalistic inquiry methods as described later in this chapter, as opposed to conventional nomothetic science—hence our use of the term *naturalistic*. Moreover, most advocates of this approach see as central the significant involvement in the evaluation of those who are participants in the endeavor being evaluated—hence the descriptor *paivicipantoriented. 50*

The evaluator portrays the different values and needs of all individuals and groups served by the program or curriculum, weighing and balancing this plurality of judgments and criteria in a largely intuitive fashion. Thus, what is judged "best" depends heavily on the values and perspectives of whichever groups or individuals are judging. By involving participants in determining the boundaries of the evaluation, evaluators serve an important educative function by creating better informed educators.

DEVELOPERS OF NATURALISTIC AND PARTICIPANT-ORIENTED EVALUATION APPROACHES AND THEIR CONTRIBUTIONS

In an important sense. Stake (1967) was the first evaluation theorist to provide significant impetus to this orientation in the Field of education. F-I's paper. .'The Countenance of Educational Evaluation," with its focus on portrayal and processing the judgments of participants, was to alter dramatically the thinking of evaluators in the next decade. Along with his later writings (Stake, 1975a. l975b, 1978, 1980), he provided conceptions and principles that have guided the evolution of this evaluation approach. Stake's early writings evidenced his growing concern over dominance of educational evaluation by parochial, objectivist, mechanistic, and stagnant conceptions and methods. Guba's (1969) discussion of the "failure of educational evaluation" provided further impetus at the time to the search for an alternative to the rationalistic approach to evaluation. Parlett and Hamilton (1976) complained that the predominant "agricultural—botanist" research paradigm was deficient for studying innovative educational programs, and they presented an alternative "illuminative evaluation" approach that followed a social anthropology paradigm. Rippey (1973) decried the insensitivity of existing evaluation approaches

to the impact of an evaluation upon the incumbents in roles within the system being evaluated; he proposed "transactional evaluation" as a more appropriate evaluation approach for systems undergoing evaluation and resultant changes. MacDonald (1974, 1976) expressed concern over existing evaluation approaches' misuses of evaluation information for questionable political purposes, opting instead for "democratic evaluation," designed to protect the rights and informational needs of the whole "community" involved. Guba and Lincoln (1981) reviewed the major approaches used in educational evaluation and rejected all except Stake's notion of responsive evaluation, which they incorporated with naturalistic inquiry to create an evaluation approach they proposed as superior to all alternatives for edueation. Patton (1975, 1978, 1980) added substantially to the literature on participant—oriented evaluation through his reports of his field—evaluation experiences. Numerous others have also suggested naturalistic or participant- oriented evaluation approaches, or methodologies that are compatible with them (for example, Kelly, 1975; MacDonald & Walker, 1977; Kemmis, 1977; Hamilton, 1976; Stenhouse, 1975; Bullock, 1982; Fetterman, 1984; and Simons, 1984, to name only a few).

Diverse as these proposals are for variants of this general evaluation approach, two threads seem to run through all of them. The first, as Wachtman (1978) notes, is disenchantment with evaluation techniques which stress a product-outcome point of view, especially at the expense of a fuller, more holistic approach which sees education as a human endeavor and admits to the complexity of the human condition. Each author

argues that instead of simplifying the issues of our humanity we should, in fact, attempt to understand ourselves and education in the context of its complexity. (Wachtman, 1978,

Second, in most of these writings, value pluralism is recognized, accommodated, and protected, even though the effort to summarize the frequently disparate judgments and preferences of such groups is left to the intuitive sagacity and communication skills of the evaluator.

Those who use the naturalistic and participant—oriented approaches to evaluation typically prepare descriptive accounts—"portrayals," as they have come to be called—of a person, classroom, school, district, project, program, activity, or some other entity around which clear boundaries have been placed. Not only is the entity richly portrayed but it is clearly positioned within the broader context in which it functions.

In addition to commonalities noted above, evaluations that use this approach generally include the following characteristics:

1. *They depend on inductit'e* ***reasoHing.*** Understanding an issue or event or process comes from grass—roots observation and discovery. Understanding emerges; it is not the end product of some preordinate inquiry plan projected before the evaluation is conducted.

2. *They use a multiplicity of data.* Understanding comes from the assimilation of data from a number of sources. Subjective and objective, qualitative and quantitative representations of the phenomena being evaluated are used.

3. *They do not follow a standard plan.* The evaluation process evolves as participants gain experience in the activity. Often the important outcome of the evaluation is a rich understanding of one specific entity with all of its idiosyncratic contextual influences, process variations, and life histories. It is important in and of itself for what it tells about the phenomena chat occurred.
4. *They record multiple rather than sinIe realities.* People see things and interpret them in different ways. No one knows everything that happens in a school. And no one perspective is accepted as *the* truth. Because only an individual can truly know what he or she has experienced, all perspectives are accepted as correct, and a central task of the evaluator is to capture these realities and portray them without sacrificing education's complexity.

Of the many authors who have proposed naturalistic and participant-oriented evaluation approaches, we have selected for further description here those whom we see as having been most influential in shaping this orientation.

Stake's Countenance Model

Stake's (1967) early analysis of the evaluation process had a major impact on evaluation thinking and laid a simple but powerful conceptual foundation for later developments in evaluation theory. He asserted that the two basic acts of evaluation are *description* and *jud,inent* (the "two countenances" of evaluation. Thus the two major activities of any formal evaluation study are full description and judgment of that which is being evaluated. To aid the evaluator in org3riizing data collection and interpretation, Stake created the evaluation framework shown in Figure 10.1.

Using thi framework, the evaluator would (1) provide background. justification, and description of the program rationale (including its need): (2) list intended antecedents (inputs, resources, existing conditions), transactions (activities, processes), and outcomes; (3) record observed antecedents, transactions. and outcomes (including observations of unintended features of each); (4) explicitly state the standards (criteria, expectations, performance of comparable programs) for judging program antecedents, transactions, and outcomes; and (5) record judgments made about the antecedent conditions, transactions, and outcomes. The evaluator would analyze information in the description matrix by looking at the congruence between intents and observations, and by looking at dependencies (contingencies) of outcomes on transactions and antecedents, and of transactions on antecedents. judgments would be made by applying standards to th descriptive data.

The countenance model thus gives evaluators a conceptual framework for thinking through the procedures of a complete evaluation.

Transactional Evaluation

Rippey (1973) used the term *transactional evaluation* to draw attention. to the effects of disruptions in an organization on incumbents in the roles in the system under—

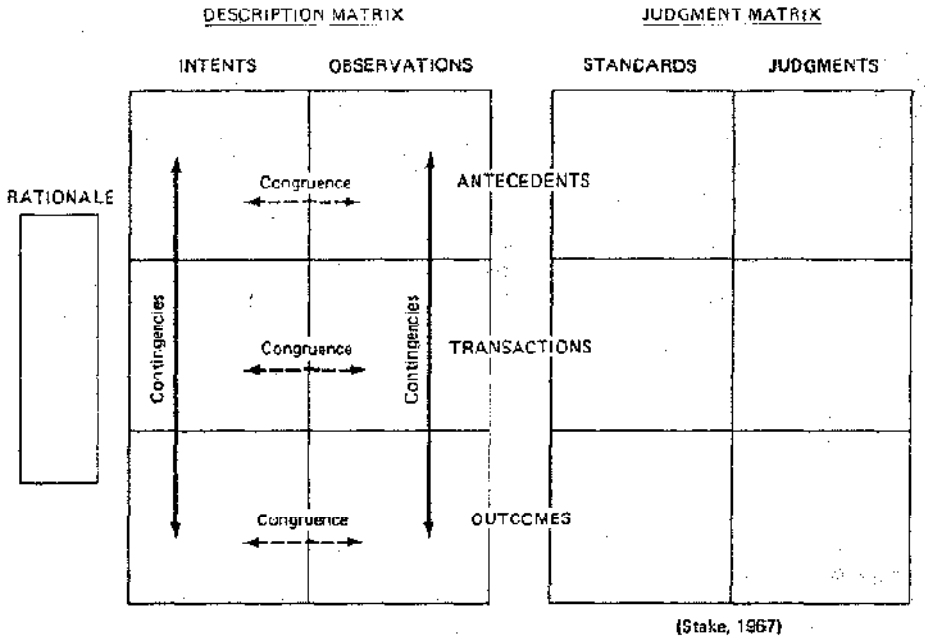


FIGURE 10.1 Stake's Layout of Statements and Data to Be Collected by the Evaluator of an Educational Program

Source: Stake, 1967.

going change. This new approach was "concerned with the system undergoing change rather than with the outcomes of the system's activity" (Rippey, 1973, p. xiii). Transactional evaluation is, therefore, a strategy for managing dysfunctions that occur within an organization in the midst of change (for instance, when a school is introducing a new program or revising an area of the curriculum). Donmoyer encapsulated it well:

> Rippey's transactional orientation emerged from the realization that programs which have been judged extremely successful in one situation can, in a different situation, be sabotaged by program participants whose values and self-interests are threatened. The transactional approach, therefore, changes evaluation's focus from a program's educational outcomes to the disequilibrium in the social system which program change inevitably brings. The approach encourages stakeholders within the social system to participate on evaluation teams and demand alterations in innovative programs which the stakeholders consider desirable. (Donmoyer, undated, p. 10)

By recognizing institutional disruptions brought about by change, transactional evaluation evidences concern with the effects on the "changers"—the staff and administrators. By attempting to alleviate these disruptions, evaluation becomes a strategy for conflict management. Rippey proposed uncovering sources of conflict and then getting proponents and opponents of the innovation involved in

developing and carrying out an evaluation plan, with the evaluation specialist providing technical assistance. Such an evaluation approach is needed, Rippey contended, because

Change, introduced in response to external needs and pressures, will have the following results:

1. It will have unexpected consequences as well as those intended.
2. It will affect the entire organization, not just the part included in the formal plan.
3. It will cause a certain amount of dislocation because of the competition for resources (including the students' time) and the shift of roles and expectations, which may

place an individual in a situation incongruent with his needs. (Rippey, 1973. p. 12)

Five phases are proposed for transactional evaluation, as follows:

1. *Initial—"trouble* spots" identified by "neutral" evaluation.
2. *Instrumentation—data* collected in meeting of various interest groups.
3. *Program development—redefinition* to reflect "group consensus" goals and values.
4. *Program monitoring—groups* agree to implement and monitor new program.
5. *Recycling—process* recycled as new conflicts emerge.

Thus, Rippey proposed using continuous evaluation of changes, by both proponents and opponents of the change, to resolve conflicts about the expected and unexpected consequences of educational innovations. The result was to move the human element in educational evaluation more directly into the spotlight. Although some may view transactional evaluation as a management-oriented evaluation approach because of its emphasis On evaluation as conflict management, or as adversary evaluation because of its inclusion of opposing pro and con points of view in planning and conducting the evaluation, we think it belongs here because of its central focus on attending to the interests of the key participants in the educational system under consideration.

Illuminative Evaluation

Parlett and Hamilton (1976) proposed an evaluation approach, which they called *illuminative evaluation,* that would involve intensive study of an educational program as a whole—its rationale and evolution, operations, achievements, and difficulties in the school context or "learning milieu." The purpose of their approach, proposed as especially applicable to small—scale programs. would be to illuminate problems, issues, and significant program features. Based on the social anthropology paradigm, and somewhat on psychiatry and sociology participant observation research, this approach grew from disenchantment with the classical experimental paradigm, which Parlett and Hamilton termed an *aricultura1-botany paradigm,* suggesting that it is a more appropriate paradigm for plants than people. Illuminative evaluation is primarily concerned with description and interpretation, not measurement and prediction. No attempt is made to manipulate or control variables, but rather to take the complex educational context, as it exists, and attempt to understand it.

The importance of studying the context of school programs, according to Parlett and *Hamilton,* is that a variety of factors influence programs in any evaluation, such as constraints (legal, *administrative,* occupational, architectural, financial); operating assumptions held by faculty (arrangement of subjects, curricula, teaching methods, grading); educators' individual characteristics (teaching style, experience, professional orientation, private goals); and students' perspectives and preoccupations. Also, the introduction of changes within the school context will set off repercussions and unusual effects. The evaluator's task is to provide a comprehensive understanding of the complex reality surrounding the program—to "illuminate" by sharpening discussion, disentangling complexities, isolating the significant from the trivial, and raising the level of sophistication characterizing debates. Although illuminative evaluation concentrates on information gathering, not decision making, **it** is expected that different groups will look to the evaluator's reports to help make difficult decisions. The illuminative evaluator does not pass judgment, but rather attempts to discover, document, and discuss what the innovation comprises and what **it** is really like to be a participant in **it.**

The process of evaluation proposed by Parlett and Hamilton has three basic stages:

1. *Observation,* to explore and become familiar with the day—co—day reality of the setting being studied.
2. *Further inquiry,* to focus the study by inquiring further on selected issues.
3. *Explanation,* to seek to explain observed patterns and cause—effect relationships.

Progressive focusing is recommended for use throughout the evaluation as a technique for refocusing and narrowing. the study, thereby allowing more concentrated attention to emerging issues.

Emphasizing classroom process, subjective information, and naturalistic inquiry, the illuminative evaluation approach depends largely on data from observations, interviews, questionnaires and tests, and documents or background sources. "Triangulative" combinations of such data are proposed to provide a more accurate portrayal of reality. The focus of this approach requires that the illuminative evaluator spend substantial periods of time in the field.

Democratic Evaluation

MacDonald (1974, 1976) delineated three tyes of evaluation studies that differed in their selection of roles, goals, audiences, techniques, and issues, as follows:

1. *Bureaucratic evaluation,* where the bureaucratic agency sponsoring the evaluation, not the evaluator, controls the evaluation information and "owns" the evaluation report.
2. *Autocratic evaluation,* where the evaluator retains ownership of the evaluation study and reports findings to the sponsoring agency and in academic journals.
3. *Democratic evaluation,* where the evaluator performs an information service to the whole community—sponsors and participants alike—with neither the evaluator nor sponsoring agency having any special claim on the findings.

Expressing a strong preference for the democratic evaluation approach, MacDonald viewed evaluation as primarily a political activity that had as its only justification the "right to know;" He argued, however, that, "People own the facts of their own lives," and he proposed that those being evaluated have the final say—veto power—over the content and release of the evaluation report.

Pluralism of values is recognized in this approach, and the range of audiences served by an evaluation report is a major criterion for judging a study's success. House (1983a) cited this evaluation approach as that which corresponds most closely to the classic liberal, individualistic approach to political pluralism:

The evaluation model that most closely corresponds to this version of liberal pluralism is MacDonald's democratic evaluation (1974). MacDonald sees the evaluator as a "broker in exchanges of information between groups," representing a range of interests, presenting information accessible to non-specialists, giving informants confidentiality and control
over the data, having no concept of information misuse, negotiating with sponsors and participants, and making no recommendations. The evaluator feeds information to audiences and lets the market work things out. Each person makes use of the information as he sees fit with the evaluator removed from interpretation. The evaluator operates on a set of procedural regulations, which control information flow. (House, 1983a. p. 61)

An interesting proposal is that the democratic evaluator should make no recommendations. MacDonald believed that evaluation reports should aspire to be nonrecommendatory "best sellers," which are widely read because of their inherent interest, nontechnical style, and usefulness and appropriateness of their information.

Responsive Evaluation

During the early 1970s, Stake began to expand his earlier (1967) writing more obviously into the realm of naturalistic and participant—oriented evaluation. Although the seeds of this explication lie in his earlier work, Stake's more recent conceptions of "responsive evaluation" (1972, 1975b, 1978, 1980) are implicitly less formal and explicitly more pluralistic than his earlier countenance model.

Responsive evaluation's central focus is in addressing the concerns and issues of a "stakeholder" audience. Stake noted that he was not proposing a new approach to evaluation, for "responsive evaluation is what people do naturally in evaluating things. They observe and react" (Stake, 1972, p. 1).52 Rather, Stake saw this approach as an attempt to develop a technology to improve and focus this natural behavior of the evaluator. Stake stressed the importance of being *responsive* to realities in the program and to the reactions, concerns, and issues of participants rather than being *preordina:e'3* with evaluation plans, relying on preconceptions and formal plans and objectives of the program. Stake defined responsive evaluation as follows:

An educational evaluation is *responsive evaluation* if it orients more directly to program
activities than to program intents; responds to audience requirements for information;

and if the different value—perspectives present are referred to in reporting the success and
failure of the program. (Stake, 1975a. p. 14)

A major reason for proposing responsive evaluation is Stake's perception that the ultimate test of an evaluation's validity is the extent to which it increases the audience's understanding of the entity that was evaluated, improved coinmunication with stakcholders is a principal goal of responsive evaluation. " The responsive approach tries to respond to the natural ways in which people assimilate information and arrive at understanding" (Stake, 1972, p. 3).

The purpose, framework, and focus of a responsive evaluation emerge from interactions with constitwcnts, and those interactions and observations result in progressive focusing on issues (similar to the progressive focusing in Parlett and Hamilton's "illuminative evaluation" described earlier). Responsive evaluators must interact continuously with members of various stakeholding groups to ascertain what information they desire and the manner in which they prefer to receive such information. Stake described the responsive evaluator's role this way:

To do a responsive evaluation, the evaluator of course does many things. He makes a plan of observations and negotiations. He arranges for various persons to observe the program. With their help he prepares for brief narratives, portrayals, product displays, graphs, etc. He finds out what is of value to his audience. He gathers expressions of worth from various individuals whose points of view differ. Of course, he checks the quality of his rccords. I-Ic gets program personnel to react to the accuracy of his portrayals. He gets authority figures to react to the importance of various findings. He gets audience members to react to the relevance of his findings. He does much of this informally, iterating, and keeping a record of action and reaction. He chooses media accessible to his audiences to increase the likelihood and fidelity of communication. He
might prepare a final written report; he might not—depending on what he and his clients have agreed on. (Stake, 1975b, p. 11)

As one might infer from the above description, responsive evaluators are relatively disinterested in formal objectives or the precision of formalized data collection; they are more likely to be at home working within the naturalistic or ethnographic paradigm, drawing heavily on qualitative techniques. Feedback to the various stakeholders is more likely to include portrayals and testimonials rather than more conventional evaluation data. Such portrayals will frequently feature descriptions of individuals in case studies based on a small sample of those affected by the program or process being evaluated. Reports to audiences will underscore the pluralism within the educational setting. A single set of recommendations is highly improbable; recommendations are more likely to be of the conditional sort where judgments about the "best" program or the "preferred" course of action will vary, depending on who is doing the judging and what criteria they use to ascertain value. Maxwell (1984) has published a rating scale that could be used to assess the quality of responsive evaluations. -

Stake (1975b) used the "clock" shown in Figure 10.2 as a mnemonic device to reflect the prominent, recurring events in a responsive evaluation:

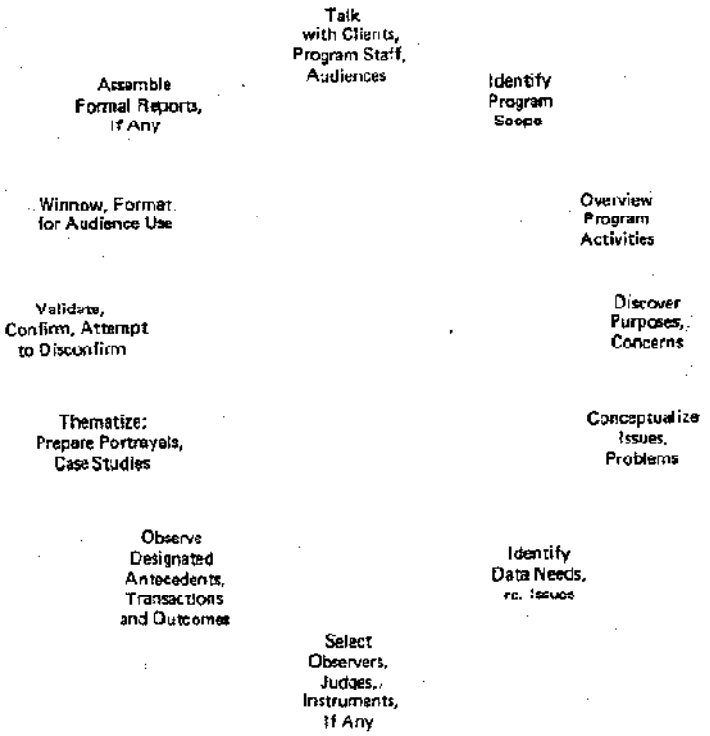


FIGURE 10.2 Prominent Events in a Responsive Evaluation

Source: Stake, 1975b, p. 19.

Although the evaluator might best begin the evaluation at twelve o'clock and proceed clockwise, Stake has emphasized that any event can follow any other event, and at any point the evaluator may want to move counterclockwise or cross-clockwise, if events warrant such flexibility. Further, many events may occur simultaneously; many will occur several times during an evaluation. The "clock" serves to remind evaluators that flexibility is an important part of using this naturalistic and participant-oriented approach.

One revealing comparison of responsive and preordinate evaluation approaches was provided by Stake's analysis of what percentage of time evaluators of each persuasion would spend on several evaluation tasks:

	Preordinate (%)	*Responsive* (%)
Identifying issues, goals	10	10
Preparing instruments	30	15
Observing the program	5	30

Administering tests, etc.	10	—
Gathering judgments	—	15
Learning client needs, etc.	—	5
Processing formal data	25	5
Preparing informal reports	—	10
Preparing formal reports	20	10

(Stake, 1975b, p. 20)

Stake (1978) also advanced acceptance of the naturalistic and participant-oriented approach to evaluation by expanding on its rationale. This approach, he said, has appeal because

1. It helps audiences for the evaluation understand the program if we pay attention to the natural way in which they understand and communicate about things.
2. Knowledge gained from experience (tacit knowledge) facilitates human understanding and extends human experience.
3. Naturalistic generalizations, which are arrived at by recognizing similarities of objects and issues in and out of context, are developed through experience. They serve to expand the way in which people come to view and understand educational programs.
4. By studying single objects, people accumulate experiences that may be used to recognize similarities in other objects. We add to existing experience and human understanding.

We have given more space to responsive evaluation than to other naturalistic and participant-oriented evaluation approaches because we agree, at least in part, with the following assessment of Guba and Lincoln (1981):

> responsive evaluation can be interpreted to include all other models. Evaluation models, as we have used that term, are differentiated on the basis of their organizers. The organizer of the responsive model is audience concerns and issues. If some audience wants to see information relating to the achievement of objectives, that is admissible within the responsive rubric. If another audience wishes to influence or service decisions, assess general effects, or elicit critical judgments, that too can be provided for within the responsive model. The responsive model can accommodate any other organizer, while other models can accommodate only the organizer on which they are based. The resulting flexibility gives the responsive model power beyond that of any of its competitors. (Guba & Lincoln, 1981, p. 38)

One might question whether or not responsive evaluation, so broadly defined, may lose its uniqueness and meaning. Such a broad claim for superiority, misunderstood, could also result in less able evaluators attempting to pass off inferior evaluations, which would be rejected as examples of any other evaluation approach, by labeling them as "responsive" evaluation. That prospect tempts one to narrow the definition of responsive studies to exclude atrocities that do not

deserve inclusion under any rubric. But one may as well rail against use of the term "creativity" because of the largely abortive efforts to define and measure that construct as to argue for limiting the broad perspective of responsive evaluation because those incapable of doing quality evaluation work, by any definition, may try to creep under the shelter of the broader conception Guba and Lincoln propose. In the final analysis, each evaluation must be judged by its usefulness, not its label. Used intelligently and competently, responsive evaluation techniques have great potential for enhancing the quality of any evaluation study.

Naturalistic Evaluation

In *The Flame Trees of Thika,* Elspeth Huxley (1982) makes this astute observation:

The best way to find things out is not to ask questions at all. If you fire off a question. **it** is like firing off a gun—bang **it** goes, and everything takes flight and runs for shelter. But if you sit quite still and pretend not to be looking, all the little facts will come and peck
round your feet, situations will venture forth from thickets, and intentions will creep out and sun themselves on a stone; and if you are very patient, you will see and understand a great deal more than a man with a gun does.

Huxley's words sum up the spirit of the naturalistic evaluation approach better than could any academic description. Yet **it** is important to move beyond the prosaic to try, as House has done, to understand the structure of reality underlying this approach. He labeled as "naturalistic" evaluation any evaluation that
aims at naturalistic generalization (based on the experience of the audience); is directed more at non—technical audiences like teachers or the general public; uses ordinary
language and everyday categories of events; and is based more on informal than formal logic. (House, 1983a. p. 57)

Although House and others had written of naturalistic approaches to evaluation, Guba (1978a) provided the first comprehensive discussion of the merits, of introducing naturalistic methods into educational evaluation. He differentiated between naturalistic inquiry, rooted in cthnography and phenomenology. and "conventional" inquiry, based on the positivist, experimental paradigm. He not only outlined several reasons for preferring naturalistic inquiry but also analyzed major methodological problems confronting naturalistic inquirers. His monograph contributed greatly toward formulation of naturalistic evaluation methodology.

The most significant work in this area, however, is the later work of Guba and Lincoln (1981), which carefully linked naturalistic inquiry to Stake's responsive evaluation, and then described procedures for implementing this approach. Their more recent work (Lincoln & Guba, 1985) extended their discussion of the naturalistic method of inquiry.

According to Guba and Lincoln, the major role of evaluation is one of responding to an audience's requirements for information in ways that take account of the different value perspectives of its members. By taking a naturalistic approach to evaluation, the evaluator is studying an educational activity *in situ,*

or as it occurs naturally, without constraining, manipulating, or controlling it.
Naturalistic inquiry casts the evaluator in the role of a learner, and those being studied in the role of informants who "teach" the evaluator. The dominant perspective is that of the informant, because the evaluators learn their perspectives, learn the concepts they use to describe their world, use their definitions of these concepts, learn the "folk theory" explanations, and translate their world so the evaluator and others can understand **jt.D4**
Guba and Lincoln stress that the criteria used to judge the rigor of scientific inquiry also hold for naturalistic inquiry, but require some reinterpretation. For instance, if one were concerned about the "truth" of an evaluation for particular subjects in a particular' context, the naturalistic evaluator would be concerned with *credibility* of findings rather than internal validity. Corroboration of data through cross—checking and triangulation are two methods used by the naturalistic evaluator to establish credibility.
if one were concerned with the *applicability* of an evaluation in other contexts or for other subjects, the naturalistic evaluator would look at the *fit* of the evaluation findings rather than external validity. Applicability is enhanced through the use of working hypotheses that should be tested in other contexts, and through the use of "thick description," which is a "literal description of the entity being evaluated, the circumstances under which it is used, the characteristics of the people involved in it, the nature of the community in which it is located, and the like" (Guba & Lincoln, 1981, p. 119).
If one were concerned with the *consistency* of evaluation findings (that is, whether the same finding would result if the study were repeated), the naturalistic evaluator would consider the study's *auditability* rather than reliability. By having a second team review the documentation and reasoning• underlying the evaluation, the evaluator can determine whether agreement on the findings can be reached. Halpern (1983) has developed an extensive model for auditing naturalistic inquiries.
Finally, if one were concerned about the *neutrality* of the evaluation, the naturalistic evaluator would look at the evaluation's *confirmability* rather than its objectivity. Data should be factual and confirmable. The naturalistic evaluator will require that the information generated by the evaluation can be confirmed.
The naturalistic evaluator proceeds by first identifying stakeholders. Their value positions are important, for it is their perspectives that should be reflected in the evaluation. Concerns and issues are elicited from interviews with the stakeholders and from naturalistic observations by the evaluator.
The naturalistic evaluator's data—collection task is defined by certain kinds of information that will be sought:
Descriptive information about the object of the evaluation and its context
Information responsive to concerns (documenting them, seeking causes and consequences, and identifying possible actions) "
Information responsive to issues (clarifying them, identifying potential courses of action to resolve them)

• Information about values (clarifying them, finding out about their source and degree of conviction)
• Information about standards to be used in the evaluation (identifying criteria, expectations, and needs).

Through the use of interviews, observations, nonverbal cues, documents, records, and unobtrusive measures, the naturalistic evaluator uses field notes and records as the sources of this information. Descriptions are used not only as data but also as a reporting technique.

Looking Back. The five naturalistic and participant-oriented evaluation approaches we have described above collectively reflect a new orientation in the field of educational evaluation. They are similar in important ways, yet each is unique. As Hamilton, Jenkins, King, MacDonald, and Parlett (1977) noted in reviewing a collection of actual evaluations that follow this general orientation, they display "a family resemblance, not an enclosed orthodoxy guided by a tacit uniformity of practice" (Hamilton and others, 1977, p. 235). Hamilton (1977), however, provided a useful general description of "pluralist evaluation models" that serves to summarize the five we have discussed:

In practical terms, pluralist evaluation models. . . can be characterized in the following manner. Compared with the classic models, they tend to be more extensive (not necessarily centered on numerical data), more naturalistic (based on program activity rather than program intent), and more adaptable (not constrained by experimental or preordinate designs). In turn, they are likely to be sensitive to the different values of program participants, to endorse empirical methods which incorporate ethnographic fieldwork, to develop feedback materials which are couched in the natural language of the
recipients, and to shift the locus of formal judgment from the evaluator to the participants. (Hamilton, 1977, p. 339)

HOW NATURALISTIC AND PARTICIPANT-ORIENTED EVALUATION APPROACHES HAVE BEEN USED

In one sense, given the breadth of this general evaluation approach, one could almost include as examples any educational evaluation that has used ethnography, case studies, storytelling, qualitative techniques, and the like. We will resist that temptation, recognizing that many studies may use some of the apparatus of participant-oriented or naturalistic evaluation without being good examples of this approach. Rather, we would point to a few examples that reflect conscious efforts to follow this evaluation approach.

Rippey (1973) has described how the concept of transactional evaluation has been used to aid in the process of change in different types of organizations. Likewise, Parlett and Dearden (1977) provide examples of the use of illuminative evaluation in evaluation of higher-education programs.

The arts in education was the focus of an extensive evaluation project using the responsive evaluation approach (Stake, l975a). In this project, and an earlier one that evaluated a program for talented youth (Stake & Gjerde, 1974)., responsive

evaluation procedures were used to address issues of immediate interest to evaluation audiences.

Malcolm and Welch (1981) provide an example of a naturalistic case study of a Catholic junior college in Minneapolis, Minnesota. Of particular interest are the authors' personal reactions to such topics as the evaluators' preparations, note-taking in the field, phases in planning for the study, interviewing, data analysis and report writing, and final editing and validation.

The use of descriptive, naturalistic case studies to report on the actual use of a new educational program is well illustrated in reports distributed by the Agency for Instructional Television (Sanders & Sonnad, 1982). Other uses of the naturalis-. tic and participant—oriented approach to evaluation have been reported by Wolcott (1976), Wolf and Tymitz (1977), Patton (1980), Guba and Lincoln (1981), Welch (1981), Spindler (1982), Hébert (1986), and Williams (1986a).

STRENGTHS AND LIMITATIONS OF NATURALISTIC AND PARTICIPANT-ORIENTED EVALUATION APPROACHES

Introduction of evaluations using this approach has prompted more acrimonious debate than almost any development in educational evaluation within the last two decades. Critics of this approach discount **it** as hopelessly "soft—headed" and argue that few if any educational evaluators are either virtuous or intellectually agile enough to wield masterfully the seductively simple yet slippery and subtle tools that this approach requires. Champions of pluralistic, responsive approaches reply that they can be readily used by any sensitive individual and that they are infinitely richer and more powerful than other approaches and, indeed, can subsume them, because they are flexible and do not preclude the use of other approaches within them, should that be desired by the evaluator's sponsor. Our intent here is not to add to what we see as a largely unproductive, divisive debate, but rather to summarize briefly some of the pros and cons thus far advanced for this approach.

Few would argue against the claim that naturalistic and participant—oriented evaluation has emphasized the human element in evaluation, it directs the attention of the evaluator to the needs of those for whom. an evaluation is being done, and it stresses the importance of a broad scope—looking at education from different viewpoints. Those who use this approach view education as a complex human undertaking and attempt to reflect that complexity as accurately as possible so that others may learn from **it.** The potential for gaining new insights and usable new theories in education from this approach stands among its greatest strengths. Other advantages of this method are its flexibility, attention to contextual variables, and encouragement of multiple data-collection techniques designed to provide a view of less tangible but crucial aspects of human and organizational behavior. In addition, this approach can provide rich and persuasive information that is credible to audiences who see it as reflecting genuine understanding of the inner workings and intricacies of the program.

As with other approaches to evaluation, the strengths of this approach may also

prove to be its limitations. Attempts to simplify the evaluation process have proven popular and effective in the past, as evidenced by the 50-year dominance of the objectives-oriented evaluation method. Thus, an approach that stresses complexity rather than simplicity may ultimately prove more popular with theorists than with practitioners, however sound it may be on other grounds.

Critics of this approach have found its subjectivity a serious limitation, even though such arguments could be mounted against every other approach to evaluation (as noted in lengthy discussions of this issue by Cuba, 1978a, and Guba & Lincoln, 1981). Because of their reliance on human observation and individual perspective, and their tendency to minimize the importance of instrumentation and group data, advocates of this approach have been criticized for "loose and unsubstantiated" evaluations. Sadler (1981) discusses 'intuitive data processing as a potential source of bias in naturalistic evaluations. Walker (1974) notes that ethnographic field work can take so much time to complete that the situation often changes or the administrator has to make a decision before the evaluation findings are available. Crittenden (1978) complains that excluding judgment from the evaluator's role makes some participant—oriented approaches nonevaluarive. Further, failure to suggest ways of weighing or combining individual standards into overall judgments about the program is viewed as making pluralistic. participant-oriented evaluation difficult for all but the most sensitive and skilled evaluators. Parlett and Hamilton (1976) concede that dependence on open-ended techniques and progressive focusing in this approach make evaluator partiality a potential problem?' Although supportive f using naturalistic techniques when the evaluation needs warrant. Williams (l986b) notes that compromises in evaluation standards are usually necessary when conducting naturalistic inquiry.

The cost of using the naturalistic and participant-oriented approach to evaluation has been viewed by some as a serious limitation, especially during times of tight budgets. This approach can be labor-intensive, often requiring full—time presence of the evaluator in the field over an extended period. The time it takes to prepare field notes and reports using this approach is at least as long as it takes for the initial observations. Some commentators (for example, Lewy, 1977) disagree. asserting that responsive—evaluation approaches take less time than other evaluation approaches, but the weight of opinion still suggests that resources (personnel. time, funds) must be considered a nontrivial limitation to wide—scale application of this approach, especially in large educational programs.

The labor intensity of naturalistic and participant—oriented approaches to evaluation limits the number of cases that can be studied intensively. Consequently, it is critical that cases be selected carefully and, even then, that conclusions not be extended beyond what those cases will allow. On the whole, evaluators using this approach are well-advised to be cautious in making interpretations and drawing conclusions. Most results might best be considered contextually anchored facts on which to base—and then test—tentative generalizations.

APPLICATION EXERCISE

As newly appointed director of Co—curricular Activities for the John F. Kennedy High School, you decide to conduct an evaluation of the co-curricular program in the school. The most current information about the program is found in the faculty handbook, published at the opening of each school year. This description reads as follows:

The John F. Kennedy High School offers a wide range of co-curricular activities for its 2,000 students. Among the various activities are clubs, intramural and varsity sports, band, choir, orchestra, and various service programs such as Red Cross. Clubs are

organized by students and assigned a faculty advisor by the Dean of Students. Meetings are scheduled on Monday—Thursday evenings and held in the cafeteria, auditorium, or gymnasium of the school. Varsity sports activities are directed by members of the

Physical Education faculty. Intramural sports are organized by home rooms and directed by a faculty member appointed by the Dean of Students. Band, choir, and orchestra are under the direction of members of the music department. Service programs are organized by students who must also find a faculty member who is willing to advise them.

You feel that this description does not provide you with sufficient insight into the program; you decide to conduct an evaluation of the current program before undertaking any modifications or restructuring of the program.

As a naturalistic and a participant—oriented evaluator, how would you proceed to plan and conduct the evaluation?

SUGGESTED READINGS

GUBA, E.G., & **LINCOLN,** Y. S. (1981). *Effective evaluation*. San Francisco: Jossey-Bass. LINCOLN, Y. S., & GUBA, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

STAKE, R. E. (1967). The countenance of educational evaluation. *Teachers College Record, 68,*

523—540. Also in B. R. WORTHEN & *j.* R. SANDERS (1973). *Educational evaluation: Theory and practice*. Belmont, CA: Wadsworth.

STAKE, R. E. (1975b). *Program evaluation, particularly responsive evaluation*. (Occasional Paper No. 5). Kalamazoo, Ml: Western Michigan University Evaluation Center.

STAKE, R. E. (1978). The case study method in social inquiry. *Educational Researcher, 7,*

5—8.